City Research Online

City, University of London Institutional Repository

Citation: Nielsen, J. P., Young, K., Mammen, E. and Byeong, U. P (2015). Asymptotics for In-Sample Density Forecasting. Annals of Statistics, 43(2), pp. 620-651. doi: 10.1214/14-AOS1288

This is the accepted version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/4960/

Link to published version: http://dx.doi.org/10.1214/14-AOS1288

Copyright: City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

Reuse: Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

City Research Online: http://openaccess.city.ac.uk/ publications@city.ac.uk

Asymptotics for In-Sample Density Forecasting

Young K. Lee[1], Enno Mammen[2], Jens P. Nielsen[3] and Byeong U. Park[4]

Kangwon National University[1], Universität Heidelberg & Higher School of Economics, Moscow[2], Cass Business School, City University London[3] and Seoul National University[4]

September 13, 2014

Abstract

This paper generalizes recent proposals of density forecasting models and it develops theory for this class of models. In density forecasting the density of observations is estimated in regions where the density is not observed. Identification of the density in such regions is guaranteed by structural assumptions on the density that allows exact extrapolation. In this paper the structural assumption is made that the density is a product of one-dimensional functions. The theory is quite general in assuming the shape of the region where the density is observed. Such models naturally arise when the time point of an observation can be written as the sum of two terms (e.g. onset and incubation period of a disease). The developed theory also allows for a multiplicative factor of seasonal effects. Seasonal effects are present in many actuarial, biostatistical, econometric and statistical studies. Smoothing estimators are proposed that are based on backfitting. Full asymptotic theory is derived for them. A practical example from the insurance business is given producing a within year budget of reported insurance claims. A small sample study supports the theoretical results.

AMS 2000 subject classifications: 62G07; 62G20

Key Words: Density estimation; kernel smoothing; backfitting; Chain Ladder.

[1]Research of Young K. Lee was supported by the NRF Grant funded by the Korea government (MEST) (No. 2010-0021396).

[2]Research of Enno Mammen was supported by the Collaborative Research Center SFB 884 "Political Economy of Reforms", financed by the German Science Foundation (DFG).

[3]Research of Jens P. Nielsen was supported by the Institute and Faculty of Actuaries, London, UK.

[4]Research of B. U. Park was supported by the NRF Grant funded by the Korea government (MEST) (No. 2010-0017437).

1 Introduction

In-sample density forecasting is in this paper defined as forecasting a structured density in regions where the density is not observed. This is possible when the density is structured in such a way that all entering components are estimable in-sample. Let us for example assume that we have one covariate X representing the start of something; it could be onset of some infection, underwriting of an insurance contract or the reporting of an insurance claim, birth of a new member of a cohort or an employee losing his job in the labour market. Let then Y represent the development or delay to some event from this starting point. It could be incubation period of some disease, development of an insurance claim, age of a cohort member or time spend looking for a new job. Then, $X+Y$ is the calendar time of the relevant event. This event is observed if and only if it has already happened until a calendar time, say t_0. The forecasting exercise is about predicting the density of future events in calendar times after t_0.

The most typical example of a structured density is a simple multiplicative form studied by Mammen, Martínez-Miranda and Nielsen (2013). The multiplicative density model assumes that X and Y are independent with smooth densities f and g. When f and g are estimated by histograms, our in-sample forecasting approach could be formulated via a parametric model. This version of in-sample density forecasting is omnipresent in academic studies as well as in business forecasting, see Martínez-Miranda, Nielsen, Sperlich, Verrall (2013) for more details and references in insurance and in statistics of cohort models. Extensions of such parametric histogram type of models can often be understood as structured density models modelled via histograms. A structured density is defined as a known function of lower-dimensional unknown underlying functions, see Mammen and Nielsen (2003) for a formal definition of generalised structured models. Under the assumption that the model is true, our forecasts do not extrapolate any parameters or time series into the future. We therefore call our methodology "in-sample density forecasting": a structured density estimator forecasting the future without further assumptions or approximate extrapolations.

Our model is related to deconvolution, but there are two major differences. First, in our

model one observes not only $X+Y$ but also the summands X and Y. Secondly, X and Y are only observed if their sum lies in a certain set, e.g., in an interval $(0, t_0]$. This destroys independence of X and Y and makes the estimation problem be an inverse problem. We will see below that the first difference leads to rates of convergence that coincide with rates for the estimation of one-dimensional functions in the classical nonparametric regression and density settings. The reason is that our model consists in a well-posed inverse problem. In contrast, deconvolution is an ill-posed inverse problem and allows only poorer rates of convergence.

This paper adds three new contributions to the literature on in-sample density forecasting. First of all, we define smoothing estimators based on backfitting and we develop a complete asymptotic distribution theory for these estimators. Secondly, we allow for a general class of regions for which the density is observed. The leading example is a triangle. A triangle arises in the above examples where the sum of two covariates is bounded by calendar time. The theoretical discussion in Mammen, Martínez-Miranda and Nielsen (2013) were restricted to this case. But there exist many other important support sets, see e.g. Kuang, Nielsen and Nielsen (2008) for a detailed discussion. Thirdly, we generalize the forecasting model by modelling a seasonal component. This is done by introducing an additional multiplicative seasonal factor into the model. Then we have three one-dimensional density functions that enter the model and that can be estimated in sample. Seasonal effects are omnipresent: onset of some disease could be more likely in the winter than in the summer; new jobs might be less likely during the summer or they may depend on the business cycle; more auto insurance claims are reported during the winter, but they might be bigger on average in the summer; cold winters or hot summers affect mortality. When a study is running over a few years only and one or two of those years are not fully observed, data might be too sparse to leave these two years out of the study. Leaving them in might however generate bias. The inclusion of seasonality in this paper solves this type of problems and allow us in general to do well when years are not fully observed. An illustration producing a within-year budget of insurance claims is given in the application section.

Classical actuarial methodology does not include seasonal effects. Budgets are nor-

mally carried out manually by highly paid actuaries. The automatic adjustment of seasonal effects offered by this paper is therefore potentially cost saving. Insurance companies currently use the classical chain ladder technique when forecasting future claims. Classical chain ladder has recently been identified as being the above mentioned multiplicative histogram in-sample forecasting approach, see Martínez-Miranda, Nielsen, Sperlich, Verrall (2013). The seasonal adjustment suggested in this paper is therefore directly implementable to working routines and processes used by today's non-life insurance companies.

Recent updates of classical chain ladder include Kuang, Nielsen and Nielsen (2009), Verrall, Nielsen and Jessen (2010), Martínez-Miranda, Nielsen, Nielsen and Verrall (2011) and Martínez-Miranda, Nielsen and Verrall (2012). These papers re-interpreted classical chain ladder in modern mathematical statistical terms. The generalised structured nonparametric model of this paper is a multiplicative density with three effects. The third seasonal effect is a function of the covariates of the first two effects. Estimation is carried out by projecting an unstructured local linear density estimator (Nielsen, 1999) down on the structure of interest. The seasonal addition to the multiplicative density model of Mammen, Martínez-Miranda and Nielsen (2013) is still a generalised additive structure, a simple special case of generalised structured models. Generalised structured models have historically been more studied in regression than in density estimation. Future developments of our in-sample density approach will therefore naturally be related to fundamental regression models, see Linton and Nielsen (1995), Nielsen and Linton (1998), Opsomer and Ruppert (1997), Mammen, Linton and Nielsen (1999), Jiang, Fan and Fan (2010), Mammen and Park (2005, 2006), Nielsen and Sperlich (2005), Mammen and Nielsen (2003), Yu, Park and Mammen (2008), Lee, Mammen and Park (2010, 2012, 2013), Zhang, Park and Wang (2013), among others.

The paper is structured as follows. Section 2 describes our structured in-sample density forecasting model, and show that the model is identifiable (estimable) under weak conditions. Section 3 explains a new approach to the estimation of the model. Here, it is assumed that the data are observed in continuous time and non-parametric smoothing methods are applied. Section 4 contains the theoretical properties of our method and Section 5 considers numerical examples and discusses the performance of the new approach.

The Appendix contains technical details.

2 The Model

We observe a random sample $\{(X_i, Y_i) : 1 \leq i \leq n\}$ from a density f supported on a subset $\mathcal{I}$ of a rectangle $[0,1]^2$. The density $f(x,y)$ of (X_i, Y_i) is a multiplicative function of three univariate components, where the first two are a function of the coordinate x and y, respectively, and the third is a function of the sum of the two coordinates, $x+y$, and is periodic. Specifically, we consider the following multiplicative model:

$$f(x,y) = f_1(x) f_2(y) f_3(\mathrm{m}_J(x+y)), \quad (x,y) \in \mathcal{I}, \tag{2.1}$$

where $\mathrm{m}_J(t) = J\mathrm{mod}_J(t)$, $\mathrm{mod}_J(t) = t$ modulo $1/J$ for some $J > 0$, i.e., $m_J(t) = J(t - l/J)$ for $l/J \leq t < (l+1)/J$, $j = 0, 1, 2 \ldots$. Here, f_j are unknown nonnegative functions supported and bounded away from zero on their supports. We note that $m_J(t)$ always takes values in $[0,1)$ as t varies on $\mathbb{R}^+$, and that the third component $f_3(\mathrm{m}_J(\cdot))$ is a periodic function with period J^{-1}.

We will prove the identifiability of the functions f_1, f_2 and f_3 under the constraints that $\int_0^1 f_1(x)\,dx = \int_0^1 f_2(y)\,dy = 1$. We will do this for two scenarios. In the first case we assume that f_1, f_2 and f_3 are smooth functions. Then identification follows by a simple argument. Our second result does not make use of smoothness conditions of the component functions. It only requires conditions on the shape of the set $\mathcal{I}$. The second result is important for an understanding of our estimation procedure that is based on a projection onto the model (2.1) without using a smoothing procedure for the component functions.

Our first identifiability result makes use of the following conditions:

(A1) The projections of the set $\mathcal{I}$ onto the x- and y-axis equal $[0,1]$.

(A2) For every $z \in [0,1)$ there exists (x,y) in the interior of $\mathcal{I}$ with $m_J(x+y) = z$. Furthermore, for every $x, y \in (0,1)$ there exist x' and y' with (x,y') and (x',y) in the interior of $\mathcal{I}$.

(A3) The functions f_1, f_2, f_3 are bounded away from zero and infinity on their supports.

(A4) The functions f_1 and f_2 are differentiable on $[0,1]$. The function f_3 is twice differentiable on $[0,1)$.

(A5) There exist sequences $x_0 = 0 < x_1 < ... < x_k = 1$ and $y_0 = 1 > y_1 > ... > y_k = 0$ with $(x, y_j) \in \mathcal{I}$ for $x_j \leq x \leq x_{j+1}$.

THEOREM 1 *Assume that model (2.1) holds with (A1)–(A5). Then the functions f_1, f_2, f_3 are identifiable.*

REMARK 1 *Let $T = \max\{x + y : (x, y) \in \mathcal{I}\}$. We note that the functions f_j are not identifiable in case $J < 1/T$. To see this, we take $f_1(u) = f_2(u) = c_1 e^u$, $f_3(u) = e^u$ with the constant $c_1 > 0$ chosen for $f_1 = f_2$ to satisfy the constraint $\int_0^1 f_j(u)\, du = 1$. Consider also $g_1(u) = g_2(u) = c_2 e^{(J+1)u}$, $g_3(u) = c_1^2/c_2^2$ with the constants $c_2 > 0$ chosen for $g_1 = g_2$ to satisfy the constraint $\int_0^1 g_j(u)\, du = 1$. In case $J < 1/T$, we have $m_J(x+y) = J(x+y)$ for all $(x, y) \in \mathcal{I}$. This implies that (f_1, f_2, f_3) and (g_1, g_2, g_3) give the same multiplicative density. In fact, if $J < 1/T$, then the assumption (A2) is not fulfilled.*

We now come to our second identifiability result that does not require smoothness conditions for the functions f_1, f_2 and f_3. This makes use of the following conditions on the shape of the support set $\mathcal{I}$. To introduce conditions on the support set $\mathcal{I}$, we let $I_1(y) = \{x : (x, y) \in \mathcal{I}\}$, $I_2(x) = \{y : (x, y) \in \mathcal{I}\}$ and $I_{3l}(z) = \{x \in [0,1] : (x, (z + l)/J - x) \in \mathcal{I}\}$. Below, we assume that these sets change smoothly as y, x and z, respectively, move. Here, $A \triangle B$ denotes the symmetric difference of two sets A and B in $\mathbb{R}$, and $\text{mes}(A)$ the Lebesgue measure of a set $A \subset \mathbb{R}$. Recall the definition $T = \max\{x + y : (x, y) \in \mathcal{I}\}$, and with this define $L(J)$ be the largest integer that is less than or equal to TJ.

(A6) For $j \in \{1, 2, 3\}$ there exist partitions $0 = a_0^j < ... < a_{L_j}^j = 1$ of $[0,1]$ and a function $\kappa : [0,1] \to \mathbb{R}^+$ with $\kappa(x) \to 0$ for $x \to 0$ such that (i) for all $u_1, u_2 \in (a_{l-1}^j, a_l^j)$, $\text{mes}[I_j(u_1) \triangle I_j(u_2)] \leq \kappa(|u_1 - u_2|)$, $l = 1, ..., L_j; j = 1, 2$; (ii) for all $u_1, u_2 \in (a_{l-1}^3, a_l^3)$, $\sum_{k=0}^{L(J)} \text{mes}[I_{3k}(u_1) \triangle I_{3k}(u_2)] \leq \kappa(|u_1 - u_2|), l = 1, ..., L_3$. Furthermore, it holds that $\text{mes}(I_2(x)) > 0$, $\text{mes}(I_1(y)) > 0$ and $\sum_{l=0}^{L(J)} \text{mes}[I_{3l}(z)] > 0$ for $x, y \in (0,1)$ and for $z \in [0,1)$.

Assumption (A6) will be used to prove the continuity of some relevant functions that appear in the technical arguments. The continuity of a function γ implies that $\gamma(x) = 0$ for all x if it is zero almost all x. The assumption allows a finite number of *jumps* in $I_j(u)$ for $j = 1, 2$ and $I_{3k}(u)$ as u moves. For example, suppose that $\mathcal{I} = \{(x, y) : 0 \leq x \leq 1, 0 \leq y \leq 1, x + y \leq 5/4\}$ and $J = 2$. In this case, $L(J) = 2$, and for $k = 0, 1$ we have $I_{3k}(z) = [0, (z + k)/2]$ for all $z \in [0, 1)$, so that I_{3k} changes smoothly as z varies on $[0, 1)$. However, for $k = 2$ we get that $I_{3k}(z) = [z/2, 1]$ for $z \in [0, 1/2]$ and $I_{3k}(z)$ is empty for $z \in (1/2, 1)$, thus it changes drastically at $z = 1/2$. In fact, $\lim_{h\to 0} \sum_{k=0}^{L(J)} \text{mes}[I_{3k}(z + h) \triangle, I_{3k}(z - h)] \neq 0$ for $z = 1/2$. We note that in this case Assumption (A6) holds if we split $[0, 1)$ into two partitions, $[0, 1/2)$ and $(1/2, 1)$.

The assumptions (A1), (A2), (A5) and (A6) accommodate a variety of sets $\mathcal{I}$ that arise in real applications. Figure 1 depicts some realistic examples of the set $\mathcal{I}$ that satisfy the assumptions. In particular, those sets of the type in the panels (c) and (e) satisfy (A2) and (A6) if the maximal vertical or horizontal thickness of the stripe is larger than the period $1/J$ of the third component function $f_3(m_J(\cdot))$. In the interpretation of the examples in Figure 1, we follow the equivalent discussion from Keiding(1990) and Kuang et al.(2008). The triangle in Figure 1a is typical for insurance or mortality when none of the underwriting years or cohorts are fully run-off. The standard actuarial term "fully run-off" means that all events from that underwriting year or cohort have been observed. In almost all practical cases of estimating outstanding liabilities, actuaries stick to the triangle format leaving out fully run-off underwriting years. While the triangle also appears in mortality studies, it is common here to leave the fully run-off cohorts in the study resulting in the support shape given in Figure 1b. The support in Figure 1c arises when the data analyst only considers observations from the most recent calendar years. While this approach is omnipresent in practical actuarial science, there is no formal theory or mathematical models behind these procedures in the actuarial literature. This paper is therefore an important step towards formalising mathematically actuarial practise while at the same time improving it. The support given in Figure 1d and Figure 1e arises when there is a known time transformation such that time is running at another pace for different underwriting years or cohort years. While this type of time transformations are

well known in mortality studies often coined as versions of accelerated failure time models. Time transformations are also well known in actuarial science coined as operational time. However, the academic literature of actuarial science is still struggling to find a formal definition of what operational time is. This paper offers one potential solution to this outstanding and important issue. The last Figure 1f is included to give an impression of the generality of support structures one could deal with inside our model approach. Data is missing in the beginning and end of the delay period, but the model is still valid and in-sample forecasts can be constructed.

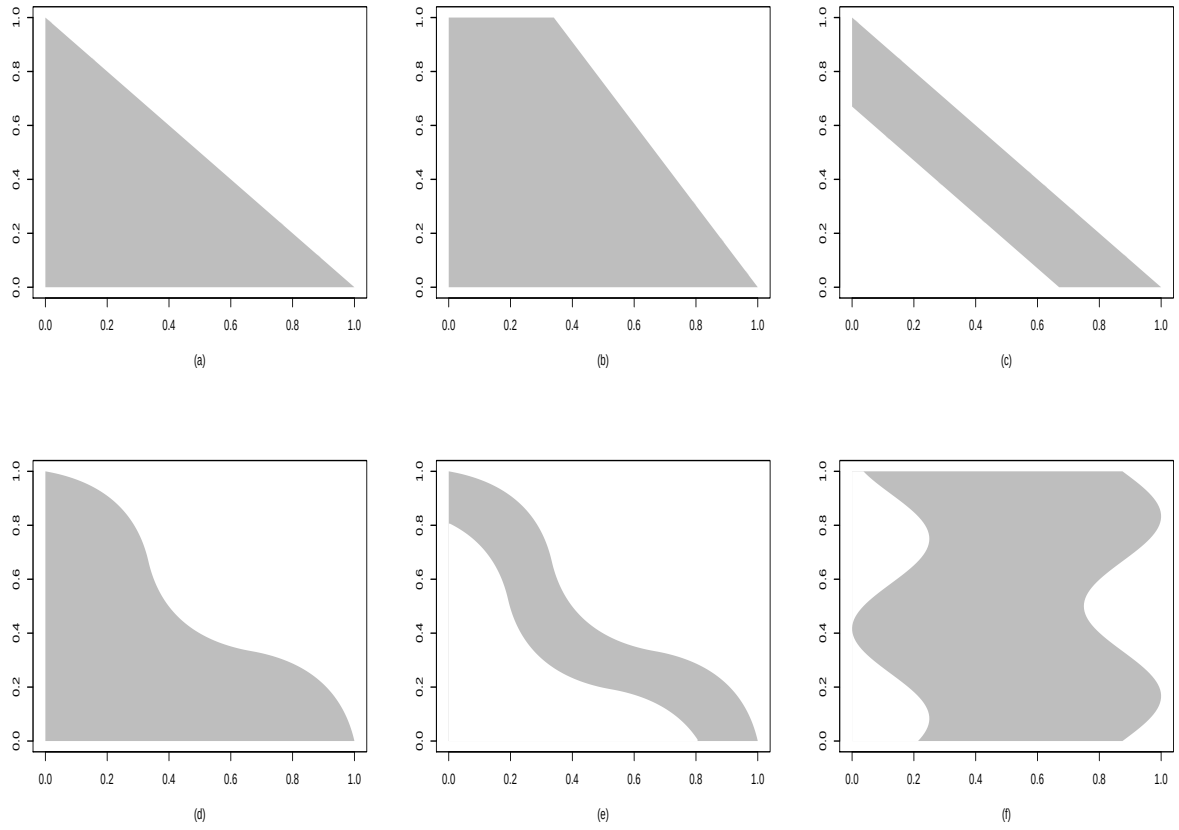

Figure 1: *Shapes of possible support sets. The horizontal axis indicates the onset (X) and the vertical the development (Y).*

The model (2.1) has taken structured density forecasting into a new territory by leaving

the simple multiplicative model. If f_3 above was constant (and therefore not in the model) then our model reduces to the simple multiplicative model analysed in Martínez-Miranda, Nielsen, Sperlich and Verrall (2013) and Mammen, Martínez-Miranda and Nielsen (2013). These two papers point out that the simple multiplicative density forecasting model is a continuous version of a widely used parametric approach corresponding to a structured histogram version of in-sample density forecasting based on the simple multiplicative model. The in-sample density forecasting model under investigation in this paper generalizes the simple multiplicative approach in an intuitive and simple way including seasonal effects.

In the following theorem, we show that, if there are two multiplicative representations of the joint density f that agree on almost all points in $\mathcal{I}$, then the component functions also agree on almost all points in $[0,1]$. We will use this result later in the asymptotic analysis of our estimation procedure.

THEOREM 2 *Assume that model (2.1) holds with (A1)–(A3), (A5), (A6). Suppose that* (g_1, g_2, g_3) *is a tuple of functions that are bounded away from zero and infinity with* $\int_0^1 g_1(x)\,dx = \int_0^1 g_2(y)\,dy = 1$. *Let* $\mu_j = \log f_j - \log g_j$. *Assume that* $\mu_1(x) + \mu_2(y) + \mu_3(m_J(x+y)) = 0$ *a.e. on* $\mathcal{I}$. *Then* $\mu_j \equiv 0$ *a.e. on* $[0,1]$.

3 Methodology

We describe the estimation method for the model (2.1). We first note that the marginal densities of X, Y and $m_J(X+Y)$ may be zero even if we assume that the joint density is bounded away from zero. For example, the marginal densities of X and Y at the point $u = 1$ are zero for the support set $\mathcal{I}$ given in Figure 1a. We estimate the multiplicative density model on a region where we observe sufficient data. This means that we exclude the points $(1,0)$ and $(0,1)$ in the estimation in the case of Figure 1a, and the point $(1,0)$ in the case of Figure 1b. Formally, for a set $S \subset \mathcal{I}$, let J_1 and J_2 denote versions of I_1 and I_2, respectively, defined by $J_1(y) = \{x : (x,y) \in S\}$ and $J_2(x) = \{y : (x,y) \in S\}$, and define $J_{3l}(z) = \{x : (x, (z+l)/J - x) \in S\}$. We take an arbitrarily small number $\delta > 0$,

and find the largest set S such that

$$\mathrm{mes}(J_2(x)) \geq \delta,\ \mathrm{mes}(J_1(y)) \geq \delta,\ \sum_{l=0}^{L(J)} \mathrm{mes}(J_{3l}(m_J(x+y))) \geq \delta \ \text{ for all } \ (x,y) \in S,$$

where $\mathrm{mes}(A)$ for a set A denotes its length. Such a set is given by $S = \{(x,y) : 0 \leq x \leq 1-\delta,\ 0 \leq y \leq 1-\delta,\ x+y \leq 1\}$ in the case of Figure 1a, and $S = \{(x,y) \in \mathcal{I} : 0 \leq x \leq 1-\delta\}$ in the case of Figure 1b, for example.

We estimate f_j on S. Let S_1 and S_2 be the projections of S onto x- and y-axis, i.e., $S_1 = \{x \in [0,1] : (x,y) \in S \text{ for some } y \in [0,1]\}$, $S_2 = \{y \in [0,1] : (x,y) \in S \text{ for some } x \in [0,1]\}$, and $S_3 = \{m_J(x+y) : (x,y) \in S\}$. In the case of Figure 1a, $S_1 = S_2 = [0, 1-\delta]$, $S_3 = [0,1)$, but in the case of Figure 1b, $S_1 = [0, 1-\delta]$, $S_2 = [0,1]$, $S_3 = [0,1)$. We put the following constraints on f_j:

$$\int_{S_1} f_1(x)\,dx = \int_{S_2} f_2(y) = 1.$$

This is only for convenience. Now, we define $f_{w,1}(x) = \int_{J_2(x)} f(x,y)\,dy$, $f_{w,2}(y) = \int_{J_1(y)} f(x,y)\,dx$ and $f_{w,3}(z) = \sum_{l=0}^{L(J)} \int_{J_{3l}(z)} f(x, (z+l)/J - x)\,dx$. Then, the model (2.1) gives the following integral equations:

$$\begin{aligned}
f_{w,1}(x) &= f_1(x) \int_{J_2(x)} f_2(y) f_3(\mathrm{m}_J(x+y))\,dy, \quad x \in S_1 \\
f_{w,2}(y) &= f_2(y) \int_{J_1(y)} f_1(x) f_3(\mathrm{m}_J(x+y))\,dx, \quad y \in S_2 \\
f_{w,3}(z) &= f_3(z) \sum_{l=0}^{L(J)} \int_{J_{3l}(z)} f_1(x) f_2((z+l)/J - x)\,dx, \quad z \in S_3.
\end{aligned} \tag{3.1}$$

We note that the marginal functions on the left hand sides of the above equations are bounded away from zero on S_j. Specifically, $\inf_{u \in S_j} f_{w,j}(u) \geq \delta \inf_{(x,y)\in\mathcal{I}} f(x,y) > 0$ so that f_j in the equations are well-defined.

Suppose that we are given a preliminary estimator of the joint density f. Call it $\hat{f}$. We estimate $f_{w,j}$ by $\hat{f}_{w,j}$ that are defined as $f_{w,j}$, respectively, with f being replaced by the preliminary estimator $\hat{f}$. Our proposed estimators of f_j, for $j = 1,2,3$, are obtained by replacing $f_{w,j}$ in the integral equations (3.1) by $\hat{f}_{w,j}$, respectively, and solving the resulting equations for the multiplicative components. Let $\vartheta = \int_S f(x,y)\,dx\,dy$ and $\hat{\vartheta}$ be

its estimator defined by $\hat{\vartheta} = n^{-1}\sum_{i=1}^{n} I[(X_i, Y_i) \in S]$. Putting the constraints

$$\int_{S_1} \hat{f}_1(x)\,dx = \int_{S_2} \hat{f}_2(y)\,dy = 1,\ \int_S \hat{f}_1(x)\hat{f}_2(y)\hat{f}_3(m_J(x+y))\,dx\,dy = \hat{\vartheta}, \tag{3.2}$$

they are given as the solution of the following backfitting equations:

$$\begin{aligned}
\hat{f}_1(x) &= \hat{\theta}_1 \cdot \frac{\hat{f}_{w,1}(x)}{\int_{J_2(x)} \hat{f}_2(y)\hat{f}_3(\mathrm{m}_J(x+y))\,dy}, \\
\hat{f}_2(y) &= \hat{\theta}_2 \cdot \frac{\hat{f}_{w,2}(y)}{\int_{J_1(y)} \hat{f}_1(x)\hat{f}_3(\mathrm{m}_J(x+y))\,dx}, \\
\hat{f}_3(z) &= \hat{\theta}_3 \cdot \frac{\hat{f}_{w,3}(z)}{\sum_{l=0}^{L(J)} \int_{J_{3l}(z)} \hat{f}_1(x)\hat{f}_2((z+l)/J - x)\,dx},
\end{aligned} \tag{3.3}$$

where $\hat{\theta}_j$ are chosen so that $\hat{f}_j$ satisfy (3.2).

The solution of (3.3) is not given explicitly. The estimates are calculated by an iterative algorithm with a starting set of function estimates $\hat{f}_1^{[0]}$ and $\hat{f}_2^{[0]}$ that satisfy the constraints (3.2). With the initial estimates, we compute $\hat{f}_3^{[0]}$ from the third equation at (3.3). Then, we update $\hat{f}_j^{[k-1]}$ consecutively for $j = 1, 2, 3$ and for $k \geq 1$ by the equations at (3.3) until convergence. Specifically, we compute at the kth cycle ($k \geq 1$) of the iteration

$$\begin{aligned}
\hat{f}_1^{[k]}(x) &= \hat{\theta}_1^{[k]} \cdot \frac{\hat{f}_{w,1}(x)}{\int_{J_2(x)} \hat{f}_2^{[k-1]}(y)\hat{f}_3^{[k-1]}(\mathrm{m}_J(x+y))\,dy}, \\
\hat{f}_2^{[k]}(y) &= \hat{\theta}_2^{[k]} \cdot \frac{\hat{f}_{w,2}(y)}{\int_{J_1(y)} \hat{f}_1^{[k]}(x)\hat{f}_3^{[k-1]}(\mathrm{m}_J(x+y))\,dx}, \\
\hat{f}_3^{[k]}(z) &= \hat{\theta}_3^{[k]} \cdot \frac{\hat{f}_{w,3}(z)}{\sum_{l=0}^{L(J)} \int_{J_{3l}(z)} \hat{f}_1^{[k]}(x)\hat{f}_2^{[k]}((z+l)/J - x)\,dx},
\end{aligned} \tag{3.4}$$

where $\hat{\theta}_j^{[k]}$ are chosen so that the resulting $\hat{f}_j^{[k]}$ satisfy (3.2).

We note that the naive two-dimensional kernel density estimator is not consistent near the boundary region, which jeopardizes the properties of the solution of the backfitting equation (3.3) at boundaries. For a preliminary estimator $\hat{f}$ of the joint density f, we take local linear estimation technique. The local linear estimator $\hat{f}$ we consider here is similar in spirit to the proposal of Cheng (1997). Let $\mathbf{a}(u, v; x, y) = (1, (u-x)/h_1, (v-y)/h_2)^\top$ and define

$$\mathbf{A}(x, y) = \int_S \mathbf{a}(u, v; x, y)\mathbf{a}(u, v; x, y)^\top h_1^{-1} h_2^{-1} K\left(\frac{u-x}{h_1}\right) K\left(\frac{v-y}{h_2}\right) du\,dv,$$

where (h_1, h_2) is the bandwidth vector and K is a symmetric univariate probability density function. Also, define

$$\hat{\mathbf{b}}(x,y) = n^{-1}\sum_{i=1}^{n} \mathbf{a}(X_i, Y_i; x, y) h_1^{-1} h_2^{-1} K\left(\frac{X_i - x}{h_1}\right) K\left(\frac{Y_i - y}{h_2}\right) W_i,$$

where $W_i = 1$ if $(X_i, Y_i) \in S$ and 0 otherwise. The local linear density estimator $\hat{f}$ we consider in this paper is defined by $\hat{\eta}_0$, where $\hat{\boldsymbol{\eta}} = (\hat{\eta}_0, \hat{\eta}_1, \hat{\eta}_2)$ is given by

$$\hat{\boldsymbol{\eta}}(x,y) = \mathbf{A}(x,y)^{-1}\hat{\mathbf{b}}(x,y). \tag{3.5}$$

It is alternatively defined as

$$\begin{aligned}\hat{\boldsymbol{\eta}}(x,y) = \arg\min_{\boldsymbol{\eta}} \lim_{b_1,b_2\to 0} \int_S \left[\hat{f}_{b_1,b_2}(u,v) - \mathbf{a}(u,v;x,y)^\top \boldsymbol{\eta}(x,y)\right]^2 \\ \times K\left(\frac{u-x}{h_1}\right) K\left(\frac{v-y}{h_2}\right) du\, dv,\end{aligned}$$

where $\hat{f}_{b_1,b_2}$ be the standard two-dimensional kernel density estimator defined by

$$\hat{f}_{b_1,b_2}(x,y) = n^{-1}\sum_{i=1}^{n} b_1^{-1} b_2^{-1} K\left(\frac{x - X_i}{b_1}\right) K\left(\frac{y - Y_i}{b_2}\right) W_i$$

for a bandwidth vector (b_1, b_2).

Before we close this section, we give two remarks. One is that, instead of integrating the two-dimensional estimator $\hat{f}$, one may estimate $f_{w,j}$ directly from the data. In particular, one may estimate $f_{w,j}$ by the one-dimensional kernel density estimators

$$\begin{aligned}\tilde{f}_{w,1}(x) &= n^{-1}h_1^{-1}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h_1}\right) W_i, \\ \tilde{f}_{w,2}(y) &= n^{-1}h_2^{-1}\sum_{i=1}^{n} K\left(\frac{Y_i - y}{h_2}\right) W_i, \\ \tilde{f}_{w,3}(z) &= n^{-1}h_3^{-1}\sum_{i=1}^{n} K\left(\frac{m_J(X_i + Y_i) - z}{h_3}\right) W_i.\end{aligned}$$

Our theory that we present in the next section is valid for this alternative estimation procedure. The other thing we would like to remark is that one may be also interested in an extension of the model (2.1) that arises when one observes a covariate $\mathbf{U}_i \in \mathbb{R}^d$ along with (X_i, Y_i). A natural extension of the model (2.1) in this case is that the conditional density

of (X, Y) given $\mathbf{U} = \mathbf{u}$ has the form $f(x, y|\mathbf{u}) = f_1(x, \mathbf{u}) f_2(y, \mathbf{u}) f_3(m_J(x+y), \mathbf{u})$, $(x, y) \in \mathcal{I}$, where the constraints (B1) now applies to $f_1(\cdot, \mathbf{z})$ and $f_2(\cdot, \mathbf{z})$ for each $\mathbf{z}$. The method and theory for this extended model are easy to derive from those we present here.

4 Theoretical Properties

Let $\mathcal{S}$ denote the space of function tuples $\mathbf{g} = (g_1, g_2, g_3)$ with square integrable univariate functions g_j in the space $L_2[0, 1]$. Define nonlinear functionals $\mathcal{F}_j$ for $1 \le j \le 3$ on $\mathcal{S}$ by

$$\begin{aligned}
\mathcal{F}_1(\mathbf{g}) &= 1 - \int_{S_1} g_1(x)\, dx, \\
\mathcal{F}_2(\mathbf{g}) &= 1 - \int_{S_2} g_2(y)\, dy, \\
\mathcal{F}_3(\mathbf{g}) &= \vartheta - \int_{S} g_1(x) g_2(y) g_3(\mathrm{m}_J(x+y))\, dx\, dy.
\end{aligned}$$

Also, define nonlinear functionals $\mathcal{F}_j$ for $4 \le j \le 6$, now on $\mathbb{R}^3 \times \mathcal{S}$, by

$$\begin{aligned}
\mathcal{F}_4(\boldsymbol{\theta}, \mathbf{g})(x) &= \int_{J_2(x)} [\theta_1 f(x, y) - g_1(x) g_2(y) g_3(\mathrm{m}_J(x+y))]\ dy, \\
\mathcal{F}_5(\boldsymbol{\theta}, \mathbf{g})(y) &= \int_{J_1(y)} [\theta_2 f(x, y) - g_1(x) g_2(y) g_3(\mathrm{m}_J(x+y))]\ dx, \\
\mathcal{F}_6(\boldsymbol{\theta}, \mathbf{g})(z) &= \sum_{l=0}^{L(J)} \int_{J_{3l}(z)} [\theta_3 f(x, (z+l)/J - x) - g_1(x) g_2((z+l)/J - x) g_3(z)]\ dx,
\end{aligned}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)^\top$. Then, we define a nonlinear operator $\mathcal{F} : \mathbb{R}^3 \times \mathcal{S} \mapsto \mathbb{R}^3 \times \mathcal{S}$ by $\mathcal{F}(\boldsymbol{\theta}, \mathbf{g})(x, y, z) = (\mathcal{F}_1(\mathbf{g}), \mathcal{F}_2(\mathbf{g}), \mathcal{F}_3(\mathbf{g}), \mathcal{F}_4(\boldsymbol{\theta}, \mathbf{g})(x), \mathcal{F}_5(\boldsymbol{\theta}, \mathbf{g})(y), \mathcal{F}_6(\boldsymbol{\theta}, \mathbf{g})(z))^\top$.

Now, we define nonlinear functionals $\hat{\mathcal{F}}_j$ for $1 \le j \le 3$ on $\mathcal{S}$ and $\hat{\mathcal{F}}_j$ for $4 \le j \le 6$ on $\mathbb{R}^3 \times \mathcal{S}$ as $\mathcal{F}_j$ in the above, with the joint density f being replaced by its estimator $\hat{f}$ and ϑ by $\hat{\vartheta}$. Let $\hat{\mathcal{F}} : \mathbb{R}^3 \times \mathcal{S} \mapsto \mathbb{R}^3 \times \mathcal{S}$ be the nonlinear operator defined by $\hat{\mathcal{F}}(\boldsymbol{\theta}, \mathbf{g})(x, y, z) = (\hat{\mathcal{F}}_1(\mathbf{g}), \hat{\mathcal{F}}_2(\mathbf{g}), \hat{\mathcal{F}}_3(\mathbf{g}), \hat{\mathcal{F}}_4(\boldsymbol{\theta}, \mathbf{g})(x), \hat{\mathcal{F}}_5(\boldsymbol{\theta}, \mathbf{g})(y), \hat{\mathcal{F}}_6(\boldsymbol{\theta}, \mathbf{g})(z))^\top$. Our estimators $\hat{\mathbf{f}} = (\hat{f}_1, \hat{f}_2, \hat{f}_3)$ along with $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)$ are given as the solution of the equation

$$\hat{\mathcal{F}}(\hat{\boldsymbol{\theta}}, \hat{\mathbf{f}}) = \mathbf{0}. \tag{4.1}$$

From the definition of the nonlinear operator $\mathcal{F}$, we also get $\mathcal{F}(\mathbf{1}, \mathbf{f}) = \mathbf{0}$, where $\mathbf{1} = (1, 1, 1)^\top$ and $\mathbf{f} = (f_1, f_2, f_3)^\top$ for the true component functions f_j.

We consider a theoretical approximation of $\hat{\mathbf{f}}$. Define a nonlinear operator by $\mathcal{G}(\boldsymbol{\theta}, \mathbf{g}) = \mathcal{F}(\mathbf{1}+\boldsymbol{\theta}, \mathbf{f} \circ (\mathbf{1}+\mathbf{g}))$, where $\mathbf{g}_1 \circ \mathbf{g}_2$ denotes the entry-wise multiplication of the two function vectors $\mathbf{g}_1$ and $\mathbf{g}_2$. Then, $\mathcal{G}(\mathbf{0}, \mathbf{0}) = \mathbf{0}$. Let $\mathcal{G}'(\mathbf{d}, \boldsymbol{\delta})$ denote the derivative of $\mathcal{G}(\boldsymbol{\theta}, \mathbf{g})$ at $(\boldsymbol{\theta}, \mathbf{g}) = (\mathbf{0}, \mathbf{0})$ to the direction $(\mathbf{d}, \boldsymbol{\delta})$. We write $\mathbf{f}_w(x, y, z) = (f_{w,1}(x), f_{w,2}(y), f_{w,3}(z))^\top$ and $\hat{\boldsymbol{\mu}}(x, y, z) = (\hat{\mu}_1(x), \hat{\mu}_2(y), \hat{\mu}_3(z))^\top$, where

$$
\begin{aligned}
\hat{\mu}_1(x) &= f_{w,1}(x)^{-1} \int_{J_2(x)} \left[\hat{f}(x, y) - f(x, y) \right] dy, \\
\hat{\mu}_2(y) &= f_{w,2}(y)^{-1} \int_{J_1(y)} \left[\hat{f}(x, y) - f(x, y) \right] dx, \\
\hat{\mu}_3(z) &= f_{w,3}(z)^{-1} \sum_{l=0}^{L(J)} \int_{J_{3l}(z)} \left[\hat{f}(x, (z+l)/J - x) - f(x, (z+l)/J - x) \right] dx.
\end{aligned}
\tag{4.2}
$$

Let $\mathcal{G}'^{-1} : \mathbb{R}^3 \times \mathcal{S} \mapsto \mathbb{R}^3 \times \mathcal{S}$ denote the inverse of $\mathcal{G}'$, whose existence we will prove in the Appendix. We define $\bar{\mathbf{f}} = (\bar{f}_1, \bar{f}_2, \bar{f}_3)$ along with $\bar{\boldsymbol{\theta}} = (\bar{\theta}_1, \bar{\theta}_2, \bar{\theta}_3)$ by

$$
\begin{pmatrix} \bar{\boldsymbol{\theta}} - \mathbf{1} \\ \\ (\bar{\mathbf{f}} - \mathbf{f})/\mathbf{f} \end{pmatrix} = \mathcal{G}'^{-1} \begin{pmatrix} \mathbf{0} \\ \\ -\mathbf{f}_w \circ \hat{\boldsymbol{\mu}} \end{pmatrix}, \tag{4.3}
$$

where $\mathbf{g}_1/\mathbf{g}_2$ denotes the entrywise division of the function $\mathbf{g}_1$ by $\mathbf{g}_2$.

It can be seen that $\boldsymbol{\delta} = (\delta_1, \delta_2, \delta_3)^\top = ((\bar{f}_1 - f_1)/f_1, (\bar{f}_2 - f_2)/f_2, (\bar{f}_3 - f_3)/f_3)^\top$ along with $\mathbf{d} = (d_1, d_2, d_3)^\top = (\bar{\theta}_1 - 1, \bar{\theta}_2 - 1, \bar{\theta}_3 - 1)^\top$ are given as the solution of the following system of integral equations.

$$
\begin{aligned}
\delta_1(x) &= d_1 + \hat{\mu}_1(x) - \int_{J_2(x)} \delta_2(y) \frac{f(x, y)}{f_{w,1}(x)} \, dy - \int_{J_2(x)} \delta_3(\mathrm{m}_J(x+y)) \frac{f(x, y)}{f_{w,1}(x)} \, dy, \quad x \in S_1 \\
\delta_2(y) &= d_2 + \hat{\mu}_2(y) - \int_{J_1(y)} \delta_1(x) \frac{f(x, y)}{f_{w,2}(y)} \, dx - \int_{J_1(y)} \delta_3(\mathrm{m}_J(x+y)) \frac{f(x, y)}{f_{w,2}(y)} \, dx, \quad y \in S_2 \\
\delta_3(z) &= d_3 + \hat{\mu}_3(z) - \sum_{l=0}^{L(J)} \int_{J_{3l}(z)} \delta_1(x) \frac{f(x, (z+l)/J - x)}{f_{w,3}(z)} \, dx \\
&\qquad - \sum_{l=0}^{L(J)} \int_{J_{3l}(z)} \delta_2((z+l)/J - x) \frac{f(x, (z+l)/J - x)}{f_{w,3}(z)} \, dx, \quad z \in S_3,
\end{aligned}
\tag{4.4}
$$

subject to the constraints

$$\begin{aligned} 0 &= \int_{S_1} f_1(x)\delta_1(x)\,dx \\ 0 &= \int_{S_2} f_2(y)\delta_2(y)\,dy \\ 0 &= \int_{S} f(x,y)\left[\delta_1(x)+\delta_2(y)+\delta_3(\mathrm{m}_J(x+y))\right]\,dx\,dy. \end{aligned} \tag{4.5}$$

In the following theorem, we show that the approximation of $\hat{\mathbf{f}}$ by $\bar{\mathbf{f}}$ is good enough. In the theorem, we assume that $\hat{f}(x,y) - f(x,y) = O_p(\varepsilon_n)$ uniformly on S for some nonnegative sequence $\{\varepsilon_n\}$ that converges to zero as n tends to infinity. For the local linear estimator $\hat{f}$ defined by (3.5) with $h_1 \sim h_2 \sim n^{-1/5}$, we have $\varepsilon_n = n^{-3/10}\sqrt{\log n}$. The theorem tells that the approximation errors of $\bar{f}_j$ for $\hat{f}_j$ are of order $O_p(n^{-3/5}\log n)$. In Theorem 4 below, we will show that $\bar{f}_j - f_j$ have magnitude of order $O_p(n^{-2/5}\sqrt{\log n})$ uniformly on S_j. This means that the first-order properties of $\hat{f}_j$ are the same as those of $\bar{f}_j$.

THEOREM 3 *Assume that the conditions of Theorem 2 hold, and that the joint density f is bounded away from zero and infinity on its support S with continuous partial derivatives on the interior of S. If $\hat{f}(x,y) - f(x,y) = O_p(\varepsilon_n)$ uniformly for $(x,y) \in S$, then it holds that $|\hat{\theta}_j - \bar{\theta}_j| = O_p(\varepsilon_n^2)$ and $\sup_{u\in S_j} |\hat{f}_j(u) - \bar{f}_j(u)| = O_p(\varepsilon_n^2)$.*

Next, we present the limit distribution of $(\bar{\mathbf{f}} - \mathbf{f})/\mathbf{f}$. In the next theorem, we assume that $h_1 \sim c_1 n^{-1/5}$ and $h_2 \sim c_2 n^{-1/5}$ for some constants c_1, $c_2 > 0$. For such constants, define

$$\tilde{f}^B(x,y) = \frac{1}{2}\int u^2 K(u)\,du \left[c_1^2 \frac{\partial^2}{\partial x^2} f(x,y) + c_2^2 \frac{\partial^2}{\partial y^2} f(x,y)\right]. \tag{4.6}$$

Also, define $\tilde{\mu}_j^B$ for $j = 1,2,3$ as $\hat{\mu}_j$ at (4.2) with the local linear estimator $\hat{f}$ being replaced by $\tilde{f}^B$. In the Appendix, we will show that the asymptotic mean of $(\bar{f}_j - f_j)/f_j$ equals $n^{-2/5}\beta_j$, where $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$ is the solution of the backfitting equation (4.4) with $\hat{\boldsymbol{\mu}}$ being replaced by $\tilde{\boldsymbol{\mu}}^B$. Let $\tilde{f}^A$ denote the centered version of the naive two-dimensional kernel density estimator. Specifically,

$$\tilde{f}^A(x,y) = n^{-1}\sum_{i=1}^{n} \left[K_{h_1}(X_i - x)K_{h_2}(Y_i - y) - E\left(K_{h_1}(X_i - x)K_{h_2}(Y_i - y)\right)\right]. \tag{4.7}$$

Here and below, we write $K_h(u) = K(u/h)/h$. Define $\tilde{\mu}_j^A$ for $j = 1, 2, 3$ as $\tilde{\mu}_j^B$ with $\tilde{f}^A$ taking the role of $\tilde{f}^B$. We will also show that the asymptotic variances of $(\bar{f}_j - f_j)/f_j$ equal those of $\tilde{\mu}_j^A$, respectively, and that they are given by $n^{-4/5}\sigma_j^2$, where

$$\begin{aligned}
\sigma_1^2(x) &= c_1^{-1} f_{w,1}(x)^{-1} \int K^2(u)\, du, \\
\sigma_2^2(y) &= c_2^{-1} f_{w,2}(y)^{-1} \int K^2(u)\, du, \\
\sigma_3^2(z) &= c_2^{-1} f_{w,3}(z)^{-1} \int [K * K(u)][K * K(c_1 u/c_2)]\, du \\
&= c_1^{-1} f_{w,3}(z)^{-1} \int [K * K(u)][K * K(c_2 u/c_1)]\, du,
\end{aligned}$$

where $K * K$ denotes the two-fold convolution of the kernel K.

In the discussion of Assumption (A6) in Section 2, we note that (A6) allows a finite number of jumps in $I_j(u)$ for $j = 1, 2$ and $I_{3l}(u)$ as u changes. These jump points are actually those where the marginal densities $f_{w,j}$ are discontinuous. At these discontinuity points the expression of the asymptotic distributions of the estimators is complicate. For this reason, we consider only those points in the partitions (a_{k-1}^j, a_k^j), $1 \le k \le L_j$, for the asymptotic distribution of $\hat{f}_j$, where a_k^j are the points that appear in Assumption (A6). We denote by $S_{j,c}$ the resulting subset of S_j after deleting all a_k^j, $1 \le k \le L_j - 1$. Note that $f_{w,j}$ is continuous on $S_{j,c}$ due to (A6). In the theorem below we also denote by S_j° the interiors of S_j, $j = 1, 2, 3$.

For the limit distribution of $\hat{f}_j$, we put an additional condition on the support set. To state the condition, let $J_2^{\circ}(u_1; h_2)$ be a subset of $J_2(u_1)$ such that $v \in J_2^{\circ}(u_1; h_2)$ if and only if $v - h_2 t \in J_2(u_1)$ for all $t \in [-1, 1]$. The set $J_2^{\circ}(u_1; h_2)$ is inside $J_2(u_1)$ at a depth h_2. In the following assumption, a_k^j and κ are the points and the function that appear in Assumption (A6).

(A7) There exist constants $C > 0$ and $\alpha > 1/2$ such that the following statements hold: (i) for any sequence of positive numbers ϵ_n, $J_2^{\circ}(u_1; C\epsilon_n^{\alpha}) \subset J_2(u_2)$ for all $u_1, u_2 \in (a_{k-1}^1, a_k^1) \cap S_1$ with $|u_1 - u_2| \le \epsilon_n$, $1 \le k \le L_1$; $J_1^{\circ}(u_1; C\epsilon_n^{\alpha}) \subset J_1(u_2)$ for all $u_1, u_2 \in (a_{k-1}^2, a_k^2) \cap S_2$ with $|u_1 - u_2| \le \epsilon_n$, $1 \le k \le L_2$; (ii) $\kappa(t) \le C|t|^{\alpha}$.

THEOREM 4 *Assume that (A7) and the conditions of Theorem 3 hold, and that the joint density f is twice partially continuously differentiable. Let the kernel K be supported on*

$[-1,1]$, *symmetric and Lipschitz continuous. Let the bandwidths* h_j *satisfy* $n^{1/5}h_j \to c_j$ *for some constants* $c_j > 0$. *Then, for fixed points* $u_j \in S_j^o \cap S_{j,c}$, *it holds that* $n^{2/5}(\bar{f}_j(u_j) - f_j(u_j))/f_j(u_j)$ *are jointly asymptotically normal with mean* $(\beta_j(u_j) : 1 \le j \le 3)$ *and variance* $\mathrm{diag}(\sigma_j(u_j) : 1 \le j \le 3)$. *Furthermore,* $(\bar{f}_j(u_j) - f_j(u_j))/f_j(u_j) = O_p(n^{-2/5}\sqrt{\log n})$ *uniformly for* $u_j \in S_j$.

REMARK 2 *In the case where the third component function* f_3 *is constant, i.e., there is no periodic component, the above theorem continue to hold for the component* f_1 *and* f_2 *without those conditions that pertain to the set* S_3 *and the function* f_3.

5 Numerical Properties

5.1 Simulation studies

We considered two densities on $\mathcal{I} = \{(x,y) : 0 \le x, y \le 1, x + y \le 1\}$. Model 1 has the components $f_1 \equiv f_2 \equiv 1$ on $[0,1]$, and $f_3(u) = c_1(\sin(2\pi u) + 3/2)$, $u \in [0,1]$, where $c_1 > 0$ is chosen to make $f(x,y) = f_1(x)f_2(y)f_3(\mathrm{m}_J(x+y))$ be a density on $\mathcal{I}$. Model 2 has $f_1(u) = 3/2 - u$, $f_2(u) = 5/4 - 3u^2/4$ and $f_3(u) = c_2(u^3 - 3u^2/2 + u/2 + 1/2)$ for some constant $c_2 > 0$. We took $J = 2$. We computed our estimates on a grid of bandwidth choice $h_1 = h_2$. For Model 1, we took $\{0.070 + 0.001 \times j : 0 \le j \le 30\}$ in the range $[0.070, 0.100]$, and for Model 2 we chose $\{0.40 + 0.02 \times j : 0 \le j \le 20\}$ in the range $[0.40, 0.80]$. In both cases, the ranges covered the optimal bandwidths. We obtained $\mathrm{MISE}_j = E\int_0^1[\hat{f}_j(u) - f_j(u)]^2du$, $\mathrm{ISB}_j = \int_0^1[E\hat{f}_j(u) - f_j(u)]^2du$ and $\mathrm{IV}_j = E\int_0^1[\hat{f}_j(u) - E\hat{f}_j(u)]^2du$, for $1 \le j \le 3$, based on 100 pseudo samples. The sample sizes were $n = 400$ and $1,000$, but only the results for $n = 400$ are reported since the lessons are the same.

Figure 2 is for Model 1. It shows the boxplots of the values of MISE_j, ISB_j and IV_j computed using the bandwidths on the grid specified above, and thus gives some indication of how sensitive our estimators are to the choice of bandwidth. The bandwidth that gave the minimal value of $\mathrm{MISE}_1 + \mathrm{MISE}_2 + \mathrm{MISE}_3$ was $h_1 = h_2 = 0.089$ in Model 1, and $h_1 = h_2 = 0.64$ in Model 2, for the sample size $n = 400$. The values of MISE_j along

with ISB_j and IV_j for these optimal bandwidths are reported in Table 1. Although our primary concern is the estimation of the component functions, it is also of interest to see how good the produced two-dimensional density estimator $\hat{f}_1(x)\hat{f}_2(y)\hat{f}_3(\mathrm{m}_J(x+y))$ behaves. For this we include in the table the values of MISE, ISB and IV of the two-dimensional estimates computed using the optimal bandwidth $h_1 = h_2 = 0.089$ in Model 1, and $h_1 = h_2 = 0.64$ in Model 2. For comparison, we also report the results for the two-dimensional local linear estimates defined at (3.5). For the local linear estimator, we used its optimal choices $h_1 = h_2 = 0.085$ in Model 1, and $h_1 = h_2 = 0.48$ in Model 2. We found that the initial local linear estimates had a large portion of mass outside $\mathcal{I}$ and thus behaved very poorly if they were not re-scaled to be integrated to one on $\mathcal{I}$. The reported values in Table 1 are for the adjusted local linear estimates. Overall, our two-dimensional estimator has better performance than the local linear estimator, especially in Model 2. Figure 3 depicts the true density of Model 1 and our two-dimensional estimate that has the median performance in terms of ISE.

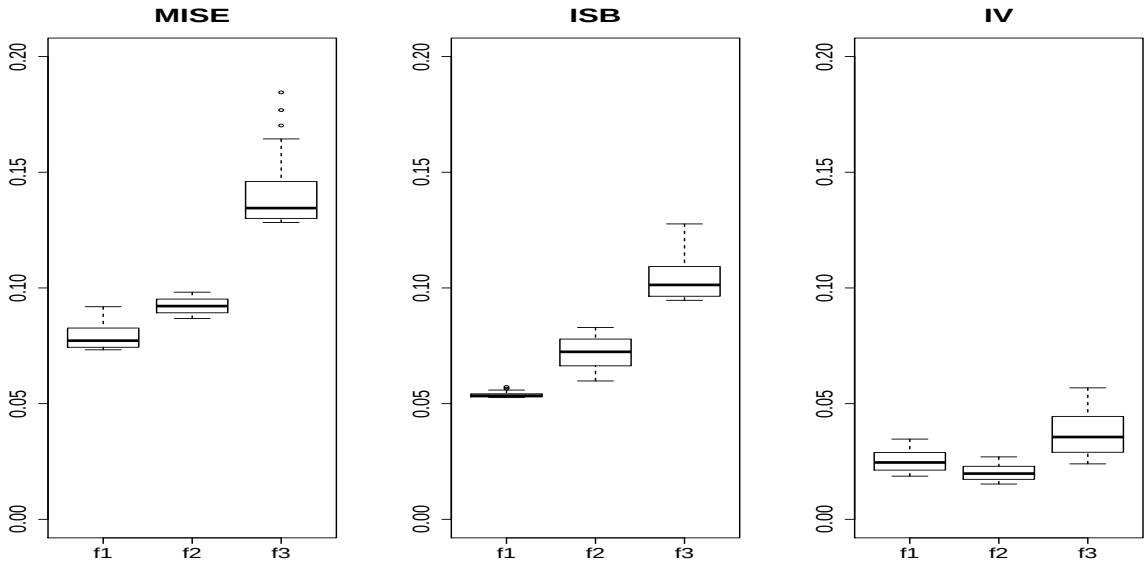

Figure 2: *Boxplots for the values of MISE, ISB and IV of our estimates f_j computed using various bandwidth choices (Model 1, $n = 400$).*

Table 1: Mean integrated squared errors (MISE), integrated squared biases (ISB) and integrated variance (IV) of the estimators.

		Component functions			Joint density	
		f_1	f_2	f_3	Our est.	Local linear
Model 1	MISE	0.0756	0.0937	0.1283	0.2493	0.2537
	ISB	0.0528	0.0752	0.0963	0.1844	0.2199
	IV	0.0228	0.0184	0.0320	0.0649	0.0338
Model 2	MISE	0.0124	0.0057	0.0130	0.0475	0.0624
	ISB	0.0120	0.0054	0.0127	0.0469	0.0607
	IV	0.0004	0.0003	0.0003	0.0006	0.0017

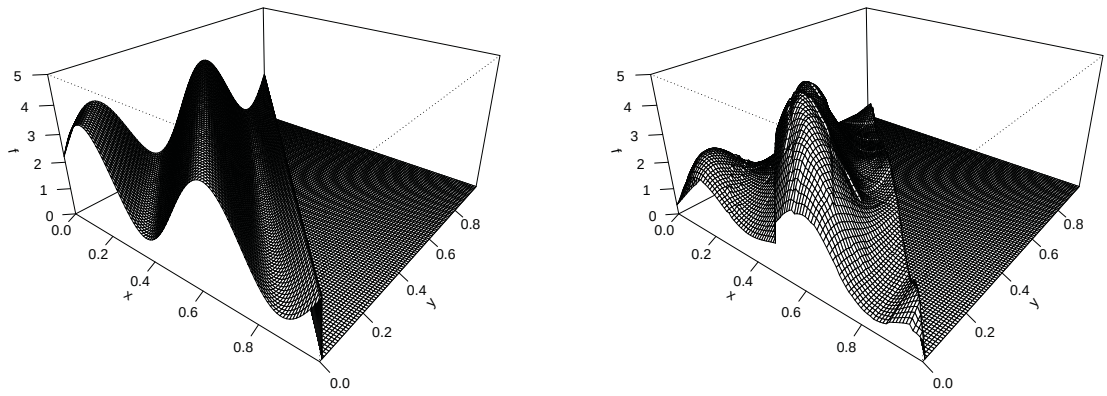

Figure 3: *The true density (left) and our estimated two-dimensional density function (right) computed from the pseudo sample that gives the median performance in terms of ISE, for Model 1 and* $n = 400$.

5.2 Data examples

The original data set we analyze in this section was collected between the year 1990 to 2011 by the major global UK based non-life insurance company RSA. The data set

– and more details about it – is publicly available via the Cass Business School web site together with the paper "Double Chain Ladder" at the Cass knowledge site. The observations were the incurred counts of large claims aggregated by months. During the 264 months 1516 large claims were made. The dataset is provided in the form of a classical run-off triangle $\{N_{kl} : 1 \leq k, l \leq 264, k+l \leq 265\}$, where N_{kl} denotes the number of large claims incurred in the kth month and reported in the $(k + l - 1)$th month i.e. with $(l - 1)$ months delay. Since the data are grouped monthly, we need pre-smoothing of the data to apply the model (2.1) that is based on data recorded over a continuous time scale. A natural way of pre-smoothing is to perturb the data by uniform random variables. Thus, we converted each claim (k, l) on the two-dimensional discrete time scale $\{(k, l) : 1 \leq k, l \leq 264, k + l \leq 265\}$, into (X, Y) on the two-dimensional continuous time scale $\mathcal{I} = \{(x, y) : 0 \leq x, y \leq 1,\ x + y \leq 1\}$, by

$$X = \frac{k - 1 + U_1}{264},\ Y = \frac{l - 1 + U_2}{264},$$

where (U_1, U_2) is a two-dimensional uniform random variate on the unit square $[0, 1]^2$. This gives a converted dataset $\{(X_i, Y_i) : 1 \leq i \leq 1516\}$. We applied to this dataset our method of estimating the structured density f of (X, Y).

Since one month corresponds to an interval with length $1/264$ on the $[0, 1]$ scale, one year is equivalent to an interval with length $12/264 = 1/22$ on the latter scale. We let the periodic component $f_3(m_J(\cdot))$ in the model (2.1) reflect a possible seasonal effect, so that we take one year in the real time to be the period of the function. This means that we let the periodic component $f_3(m_J(\cdot))$ have $1/22$ as its period, and thus take $J = 22$. For the bandwidth we took $h_1 = h_2 = 0.01$. The chosen bandwidth may be considered to be too small for the estimation of f_1 and f_2. However, we took such a small bandwidth to detect possible seasonality. Note that the bandwidth size 0.01 corresponds to $0.01 \times 12 \times 22 = 2.64$ months. We found that even with this small bandwidth the estimated curve $\hat{f}_3$ was nearly a constant function, which suggests that the large claim data do not have a seasonal effect.

To see how well our method detects a possible seasonal effect in the data, we augmented

the dataset by adding a certain level of seasonal effect as follows. We computed

$$
\begin{aligned}
N'_{kl} &= 2\,N_{kl} \quad \text{if } k+l = 12m \text{ for some } m = 1, 2, \ldots, \\
N'_{kl} &= 3\,N_{kl} \quad \text{if } k+l = 12m+1 \text{ for some } m = 1, 2, \ldots, \\
N'_{kl} &= 5\,N_{kl} \quad \text{if } k+l = 12m+2 \text{ for some } m = 0, 1, \ldots, \\
N'_{kl} &= 3\,N_{kl} \quad \text{if } k+l = 12m+3 \text{ for some } m = 0, 1, \ldots, \\
N'_{kl} &= N_{kl} \quad \text{otherwise.}
\end{aligned}
$$

Since $(k + l - 1 \text{ modulo } 12)$ is the actual month of the claims reported, the augmented dataset has added claims in November, December, January and February. The augmentation resulted in increasing the total number of claims to 2606 from 1516. The increased counts of reported claims were 252 from 126 for November, 600 from 200 for December, 455 from 91 for January, and 300 from 100 for February.

In our estimation procedure, the bandwidths h_1 and h_2 control the smoothness of the local linear estimate $\hat{f}$ along the x- and y-axis, respectively. Consequently, choosing small values for h_1 and h_2 would result in non-smooth estimates of the functions f_1 and f_2, which we observed in the pilot study with $h_1 = h_2 = 0.01$. Nevertheless, in some cases setting these bandwidths to be small, relative to the scales of X and Y, might be preferred when one needs to detect possible seasonality, as is the case with the current dataset. In our dataset the bandwidth size $1/264 = 0.0038$ on the scale of $[0, 1]$ corresponds to one month in real time. Thus, taking the bandwidths to be 0.015, for example, that corresponds to a period of four months, forces the seasonal effect to almost vanish in the estimate of f_3.

To achieve both aims of producing smooth estimates of f_1 and f_2, and of detecting possible seasonal effect, we applied to the augmented dataset a two-stage procedure that is based on our estimation method described in Section 3. In the first stage, we got a local linear estimate $\hat{f}$ with $h_1 = h_2 = 0.01$, and found an estimate of f_3 using the iteration scheme at (3.4). In the second stage, we recomputed a local linear estimate $\hat{f}$ with larger bandwidths $h_1 = h_2 = 0.05$, and found estimates of f_1 and f_2 using only the first two updating equations at (3.4) with $\hat{f}_3^{[k-1]}$ being replaced by the estimate of f_3 obtained in the first stage.

The results of applying this two-stage procedure to the augmented dataset are pre-

sented in Figure 4. Clearly, the seasonal effect of the augmented dataset was well recovered in the estimate of f_3, and at the same time smooth estimates of f_1 and f_2 were produced. The augmented data set indicate an increased number of claims in the winter time. This is clearly reflected in the estimated results, where the first part and the last part of the estimated effect is higher than the rest of the curve. Imagine the realistic situation that a non-life insurer on the first day of November has to produce budget expenses for the rest of the year. The classical multiplicative methodology is not able to reflect the two month perspective of such a budget. Therefore considerable work is being done manually in Finance and Actuarial departments of non-life insurance companies to correct for such effects. With our new seasonal correction costly manual procedures can be replaced by cost saving automatic ones eventually benefitting the prices all of us as end customers have to pay for insurance products.

Figure 5 depicts the resulting two-dimensional joint density. Notice that this two-dimensional density is clearly non-multiplicative. The seasonal correction provides a visually deviation from the multiplicative shape. Also, note that while this two-dimensional density is non-multiplicative, the nature of this deviation is not immediately clear to the eye. Whether the deviation is pure noise, a seasonal effect or some other effect is not easy to get from the full two-dimensional graph of the local linear density estimate which is also presented in Figure 5. For the local linear estimate we used $h_1 = h_2 = 0.03$. We tried other bandwidth choices such as 0.01 and 0.05, but found that the smaller one gave too rough estimate and the larger one produced too smooth a surface. Our two-dimensional density estimate therefore illustrates why research into structured densities on non-trivial supports is crucial to extract information beyond the classical and simple multiplicative one.

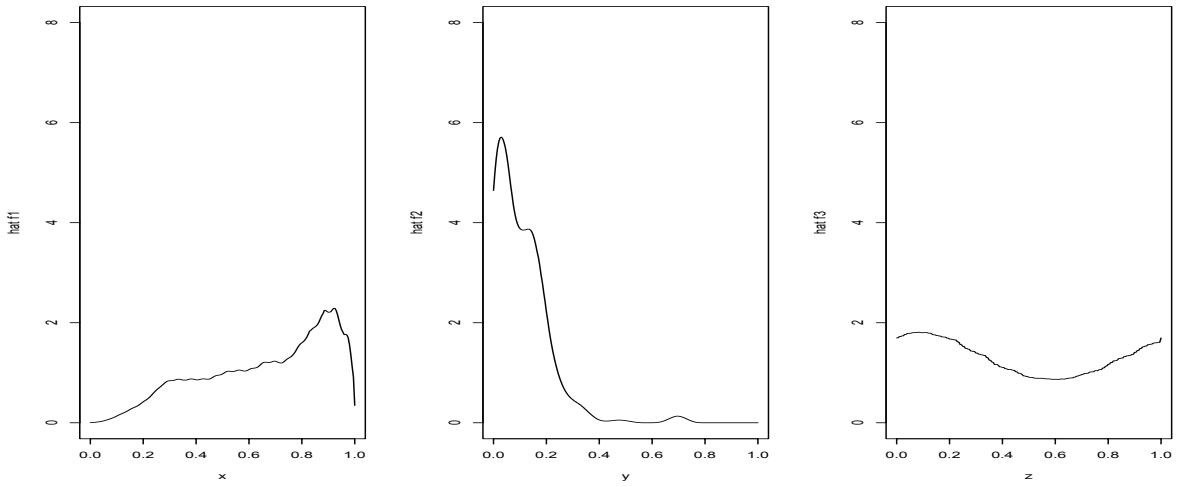

Figure 4: *Estimated curves $\hat{f}_j$ for the model (2.1) obtained by applying the two-stage procedure to the augmented large claim data.*

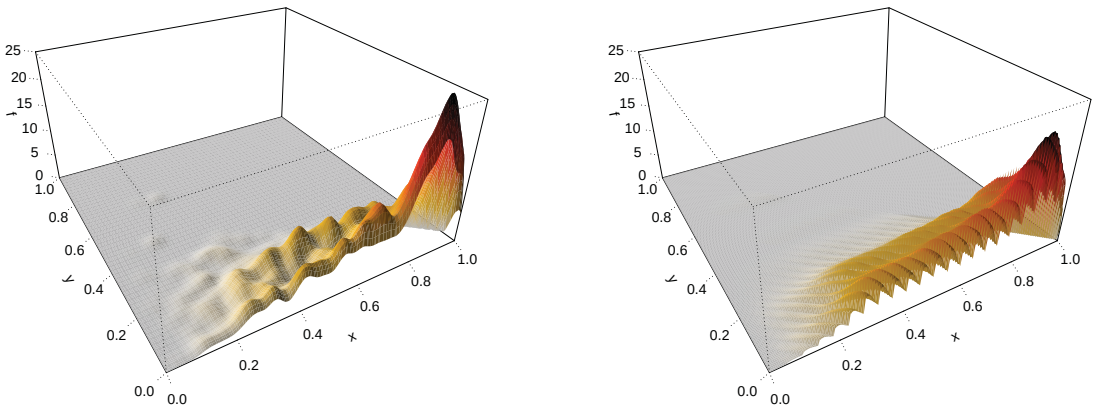

Figure 5: *Local linear joint density estimate (left) and our estimate (right) for the model (2.1) obtained by applying the two-stage procedure to the augmented large claim data*

Appendix

A.1 Proof of Theorem 1

Suppose that (g_1, g_2, g_3) is a tuple of functions that are bounded away from zero and infinity with $\int_0^1 g_1(x)\,dx = \int_0^1 g_2(y)\,dy = 1$ and

$$f(x, y) = g_1(x)g_2(y)g_3(\mathrm{m}_J(x + y)).$$

Furthermore, we assume that g_1 and g_2 are differentiable on $[0,1]$ and that g_3 is twice differentiable on $[0,1)$. For $j \in \{1,2,3\}$ define $\mu_j = \log f_j - \log g_j$. By assumption we have

$$\mu_1(x) + \mu_2(y) + \mu_3(m_J(x+y)) = 0.$$

For $z \in [0,1)$ we choose (x,y) in the interior of $\mathcal{I}$ with $m_J(x+y) = z$. Then we have that

$$0 = \frac{\partial^2}{\partial x \partial y}[\mu_1(x) + \mu_2(y) + \mu_3(m_J(x+y))] = \mu_3''(z).$$

Thus μ_3 is a linear function. Furthermore, we have that $\mu_3(0) = \mu_3(1-)$. This follows by noting that $\mu_3(0) = -\mu_1(x) - \mu_2(y)$ for $(x,y) \in \mathcal{I}$ with $m_J(x+y) = 0$. Note that $m_J(x+y) = 0$ if and only if $x + y = l/J$ for some $l \geq 1$, if (x,y) is in the interior of $\mathcal{I}$. After slightly decreasing x and y to $x + \delta_x$ and $y + \delta_y$ with small $\delta_x < 0$, $\delta_y < 0$ we have that $\mu_3(1 + J(\delta_x + \delta_y)) = -\mu_1(x+\delta_x) - \mu_2(y+\delta_y)$ since $m_J(x+y+\delta_x+\delta_y) = 1 + J(\delta_x + \delta_y)$. Thus $\mu_3(0) = \mu_3(1-)$ follows from continuity of μ_1 and μ_2. We conclude that μ_3 must be a constant function. Thus $\mu_1(x) + \mu_2(y)$ is a constant function.

From Assumption (A5) we get that $\mu_1(x)$ is constant on the intervals $[x_j, x_{j+1}]$. Because the union of these intervals is equal to $[0,1]$ we conclude that $\mu_1(x)$ is constant on $[0,1]$. Using again (A5) we get that $\mu_2(y)$ is constant on $[0,1]$. Because of the assumption that $\int_0^1 g_1(x)\,dx = \int_0^1 g_2(y)\,dy = 1$ and $\int_0^1 f_1(x)\,dx = \int_0^1 f_2(y)\,dy = 1$ we get that $f_1 = g_1$, $f_2 = g_2$ and $f_3 = g_3$. This concludes the proof.

A.2 Proof of Theorem 2

We first argue that μ_1, μ_2 and μ_3 are a.e. equal to piecewise continuous functions on $(0,1)$, with a finite number of pieces. To see that μ_1 is a.e. equal to a piecewise continuous function, we note that

$$\mu_1(x) = -\int_{I_2(x)} [\mu_2(y) + \mu_3(m_J(x+y))]\ dy/\text{mes}(I_2(x)),\ a.e.\ x \in (0,1).$$

Here, because of (A3) and (A6), the right hand side is a piecewise continuous function. Thus, μ_1 is a.e. equal to a piecewise continuous function. In abuse of notation, we now denote the piecewise continuous function by μ_1. By similar arguments one sees that

μ_2, and μ_3 are piecewise continuous functions (or more precisely a.e. equal to piecewise continuous functions). This implies that

$$\mu_1(x) + \mu_2(y) + \mu_3(m_J(x+y)) = 0 \tag{A.1}$$

for $(x, y, m_J(x+y)) \notin \{x_1, ..., x_{r_1}\} \times (0,1)^2 \cup (0,1) \times \{y_1, ..., y_{r_2}\} \times (0,1) \cup (0,1)^2 \times \{z_1, ..., z_{r_3}\}$ for some values $x_1, ..., x_{r_1}, y_1, ..., y_{r_2}, z_1, ..., z_{r_3} \in (0,1)$.

We now argue that μ_3 is continuous on $[0,1)$. To see that μ_3 is continuous at $z_0 \in [0,1)$, we choose (x_0, y_0) in the interior of $\mathcal{I}$ such that $m_J(x_0+y_0) = z_0$. This is possible because of Assumption (A2). We can choose x_0 and y_0 such that μ_1 is continuous at x_0 and μ_2 is continuous at y_0. Thus we get from (A.1) that μ_3 is continuous at z_0. Similarly one shows that μ_1 and μ_2 are continuous functions on $[0,1]$. This gives that

$$\mu_1(x) + \mu_2(y) + \mu_3(m_J(x+y)) = 0 \tag{A.2}$$

for all $x, y \in (0,1)$.

For $z_0 \in [0,1)$ we choose (x_0, y_0) in the interior of $\mathcal{I}$ with $m_J(x_0 + y_0) = z_0$. Note that for δ_x and δ_y sufficiently small we get for $z_0 \in (0,1)$ that $m_J(x_0 + \delta_x + y_0 + \delta_y) = z_0 + J(\delta_x + \delta_y)$. This gives for δ_x and δ_y sufficiently small that

$$\mu_1(x_0 + \delta_x) + \mu_2(y_0 + \delta_y) + \mu_3(z_0 + J(\delta_x + \delta_y)) = 0.$$

With δ_x, δ_y' and δ_y sufficiently small we get that

$$\mu_2(y_0 + \delta_y) + \mu_3(z_0 + J(\delta_x + \delta_y)) = \mu_2(y_0 + \delta_y') + \mu_3(z_0 + J(\delta_x + \delta_y')).$$

With the special choice $\delta_x = -\delta_y$ this gives

$$\mu_2(y_0 + \delta_y) + \mu_3(z_0) = \mu_2(y_0 + \delta_y') + \mu_3(z_0 + J(\delta_y' - \delta_y)).$$

Let γ be a function defined by $\gamma(u) = \mu_3(z_0 + Ju) - \mu_3(z_0)$. From the last two equations taking $u = \delta_x + \delta_y$ and $v = \delta_y' - \delta_y$, we get

$$\gamma(u+v) = \gamma(u) + \gamma(v)$$

for u, v sufficiently small. This implies that, with a constant c_{z_0} depending on z_0 we have $\gamma(u) = c_{z_0} u$ for u sufficiently small, see Theorem 3 of Guillot et al. (2013). Thus, we obtain

$\mu_3(z) = a_{z_0} + b_{z_0} z$ with constants a_{z_0} and b_{z_0} depending on z_0 for z in a neighborhood U_{z_0} of z_0. Because every interval $[z', z'']$ with $0 < z' < z'' < 1$ can be covered by the union of finitely many U_z's we get that for each such interval it holds that $\mu_3(z) = a_{z',z''} + b_{z',z''} z$ for $z \in [z', z'']$ with constants $a_{z',z''}$ and $b_{z',z''}$ depending on the chosen interval $[z', z'']$.

One can repeat the above arguments for $z_0 = 0$. Then we have that $m_J(x_0 + \delta_x + y_0 + \delta_y) = 1 + J(\delta_x + \delta_y)$ for $\delta_x + \delta_y < 0$ and $m_J(x_0 + \delta_x + y_0 + \delta_y) = J(\delta_x + \delta_y)$ for $\delta_x + \delta_y > 0$. Arguing as above with $\delta_x + \delta_y > 0$ and $\delta'_y - \delta_y > 0$ we get that $\mu_3(z) = a_+ + b_+ z$ for $z \in (0, z^+]$ for $z^+ > 0$ small enough with some constants a_+ and b_+. Similarly we get by choosing $\delta_x + \delta_y < 0$ and $\delta'_y - \delta_y < 0$ that $\mu_3(z) = a_- + b_- z$ for $z \in (z^-, 1)$ for $z^- < 1$ large enough with some constants a_- and b_-. Thus we get that $\mu_3(z) = a + bz$ for $z \in (0, 1)$ with some constants a and b.

Furthermore, using continuity of μ_1, μ_2 and the relation $\mu_3(m_J(x + y)) = -\mu_1(x) - \mu_2(y)$ for $z = m_J(x + y)$ with z in $(1 - \delta, 1)$ and $(0, \delta)$ with $\delta > 0$ small enough we get that $\mu_3(0) = \mu_3(1-)$. Thus we have $b = 0$ and we conclude that μ_3 is a constant function. This gives

$$\mu_1(x) + \mu_2(y) = -a$$

for all $(x, y) \in \mathcal{I}$. Now arguing as in the proof of Theorem 1 we get that $f_1 = g_1$, $f_2 = g_2$ and $f_3 = g_3$. This concludes the proof.

A.3 Proof of Theorem 3

Let $\mathcal{G}'(\boldsymbol{\theta}, \mathbf{g})(\mathbf{d}, \boldsymbol{\delta})$ denote the derivative $\mathcal{G}$, defined in Section 4, at $(\boldsymbol{\theta}, \mathbf{g})$ to the direction $(\mathbf{d}, \boldsymbol{\delta})$. We note that we write $\mathcal{G}'(\mathbf{0}, \mathbf{0})(\mathbf{d}, \boldsymbol{\delta})$ simply as $\mathcal{G}'(\mathbf{d}, \boldsymbol{\delta})$ in Section 4. We use the sup-norm $\|(\mathbf{d}, \boldsymbol{\delta})\|_\infty$ as a metric in the space $\mathbb{R}^3 \times \mathcal{S}$, defined by

$$\|(\mathbf{d}, \boldsymbol{\delta})\|_\infty = \max\Big\{|d_1|, |d_2|, |d_3|, \sup_{u \in S_1} |\delta_1(u)|, \sup_{u \in S_2} |\delta_2(u)|, \sup_{u \in S_3} |\delta_3(u)|\Big\}.$$

Define $\hat{\mathcal{G}}(\boldsymbol{\theta}, \mathbf{g}) = \hat{\mathcal{F}}(\mathbf{1} + \boldsymbol{\theta}, \mathbf{f} \circ (\mathbf{1} + \mathbf{g}))$, where $\hat{\mathcal{F}}$ is defined in Section 4, and let $\hat{\mathcal{G}}'(\boldsymbol{\theta}, \mathbf{g})$ denote the derivative of $\hat{\mathcal{G}}$ at $(\boldsymbol{\theta}, \mathbf{g})$. In the setting where $\hat{f}(x, y) - f(x, y) = O_p(\varepsilon_n)$ uniformly for $(x, y) \in \mathcal{I}$, we claim

(i) $\sup_{\|(\mathbf{d}, \boldsymbol{\delta})\|_\infty = 1} \|\hat{\mathcal{G}}'(\mathbf{0}, \mathbf{0})(\mathbf{d}, \boldsymbol{\delta}) - \mathcal{G}'(\mathbf{0}, \mathbf{0})(\mathbf{d}, \boldsymbol{\delta})\|_\infty = O_p(\varepsilon_n)$;

(ii) The operator $\mathcal{G}'(\mathbf{0},\mathbf{0})$ is invertible and has bounded inverse;

(iii) The operator $\hat{\mathcal{G}}'$ is Lipschitz continuous with probability tending to one, i.e., there exists constants r, $C > 0$ such that, with probability tending to one,

$$\sup_{\|(\mathbf{d},\boldsymbol{\delta})\|_\infty=1} \|\hat{\mathcal{G}}'(\boldsymbol{\theta}_1,\mathbf{g}_1)(\mathbf{d},\boldsymbol{\delta}) - \hat{\mathcal{G}}'(\boldsymbol{\theta}_2,\mathbf{g}_2)(\mathbf{d},\boldsymbol{\delta})\|_\infty \leq C\|(\boldsymbol{\theta}_1,\mathbf{g}_1) - (\boldsymbol{\theta}_2,\mathbf{g}_2)\|_\infty$$

for all $(\boldsymbol{\theta}_1,\mathbf{g}_1)$, $(\boldsymbol{\theta}_2,\mathbf{g}_2) \in B_r(\mathbf{0},\mathbf{0})$, where $B_r(\boldsymbol{\theta},\mathbf{g})$ is a ball with radius $r > 0$ in $\mathbb{R}^3 \times \mathcal{S}$ centered at $(\boldsymbol{\theta},\mathbf{g})$.

Theorem 3 basically follows from the above (i)–(iii). To prove the theorem using (i)–(iii), we note that Claim (ii) with the definitions of $\bar{\boldsymbol{\theta}}$ and $\bar{\mathbf{f}}$ at (4.3) gives $\bar{\boldsymbol{\theta}} - \mathbf{1} = O_p(\varepsilon_n)$ and $(\bar{\mathbf{f}} - \mathbf{f})/\mathbf{f} = O_p(\varepsilon_n)$. With (i) and (iii), this implies that

$$\sup_{\|(\mathbf{d},\boldsymbol{\delta})\|_\infty=1} \|\hat{\mathcal{G}}'(\bar{\boldsymbol{\theta}} - \mathbf{1}, (\bar{\mathbf{f}} - \mathbf{f})/\mathbf{f})(\mathbf{d},\boldsymbol{\delta}) - \mathcal{G}'(\mathbf{0},\mathbf{0})(\mathbf{d},\boldsymbol{\delta})\| = O_p(\varepsilon_n). \tag{A.3}$$

Now, from (ii) it follows that there exists a constant $C > 0$ such that the map $\hat{\mathcal{G}}'(\bar{\boldsymbol{\theta}} - \mathbf{1}, (\bar{\mathbf{f}}-\mathbf{f})/\mathbf{f})$ is invertible and $\|\hat{\mathcal{G}}'(\bar{\boldsymbol{\theta}}-\mathbf{1}, (\bar{\mathbf{f}}-\mathbf{f})/\mathbf{f})^{-1}(\mathbf{d},\boldsymbol{\delta})\|_\infty \leq C\|(\mathbf{d},\boldsymbol{\delta})\|_\infty$ with probability tending to one. Also, (iii) is valid for all $(\boldsymbol{\theta}_1,\mathbf{g}_1)$, $(\boldsymbol{\theta}_2,\mathbf{g}_2) \in B_{2r}(\bar{\boldsymbol{\theta}} - \mathbf{1}, (\bar{\mathbf{f}} - \mathbf{f})/\mathbf{f})$. Then, we can argue that the solution of the equation $\hat{\mathcal{G}}(\boldsymbol{\theta},\mathbf{g}) = \mathbf{0}$, which is $(\hat{\boldsymbol{\theta}} - \mathbf{1}, (\hat{\mathbf{f}} - \mathbf{f})/\mathbf{f})$, is within $C\alpha_n$ distance from $(\bar{\boldsymbol{\theta}} - \mathbf{1}, (\bar{\mathbf{f}} - \mathbf{f})/\mathbf{f})$, with probability tending to one, where $C > 0$ is a constant and $\alpha_n = \|\hat{\mathcal{G}}(\bar{\boldsymbol{\theta}} - \mathbf{1}, (\bar{\mathbf{f}} - \mathbf{f})/\mathbf{f})\|_\infty$. This follows from an application of Newton-Kantorovich theorem, see Deimling (1985) or Yu, Park and Mammen (2008) for a statement of the theorem and related applications. To compute α_n we note that

$$\begin{aligned}\hat{\mathcal{G}}(\bar{\boldsymbol{\theta}} - \mathbf{1}, (\bar{\mathbf{f}} - \mathbf{f})/\mathbf{f}) &= \hat{\mathcal{G}}(\mathbf{0},\mathbf{0}) + \hat{\mathcal{G}}'(\mathbf{0},\mathbf{0})(\bar{\boldsymbol{\theta}} - \mathbf{1}, (\bar{\mathbf{f}} - \mathbf{f})/\mathbf{f}) + O_p(\varepsilon_n^2) \\ &= \hat{\mathcal{G}}(\mathbf{0},\mathbf{0}) + \mathcal{G}'(\mathbf{0},\mathbf{0})(\bar{\boldsymbol{\theta}} - \mathbf{1}, (\bar{\mathbf{f}} - \mathbf{f})/\mathbf{f}) + O_p(\varepsilon_n^2).\end{aligned} \tag{A.4}$$

For the first equation of (A.4) we have used (iii) and the facts that $\bar{\boldsymbol{\theta}} - \mathbf{1} = O_p(\varepsilon_n)$ and $(\bar{\mathbf{f}} - \mathbf{f})/\mathbf{f} = O_p(\varepsilon_n)$. The second equation of (A.4) follows from the inequality

$$\|\hat{\mathcal{G}}'(\mathbf{0},\mathbf{0})(\mathbf{d},\boldsymbol{\delta}) - \mathcal{G}'(\mathbf{0},\mathbf{0})(\mathbf{d},\boldsymbol{\delta})\|_\infty \leq C \sup_{x,y\in S} |\hat{f}(x,y) - f(x,y)| \cdot \|(\mathbf{d},\boldsymbol{\delta})\|_\infty$$

for some constant $C > 0$. Now, $\hat{\mathcal{G}}(\mathbf{0},\mathbf{0}) = \hat{\mathcal{F}}(\mathbf{1},\mathbf{f}) = (\mathbf{0}^\top, (\mathbf{f}_w \circ \hat{\boldsymbol{\mu}})^\top)^\top$. From the definition (4.3), we also get $\mathcal{G}'(\mathbf{0},\mathbf{0})(\bar{\boldsymbol{\theta}} - \mathbf{1}, (\bar{\mathbf{f}} - \mathbf{f})/\mathbf{f}) = (\mathbf{0}^\top, -(\mathbf{f}_w \circ \hat{\boldsymbol{\mu}})^\top)^\top$. This proves $\alpha_n = O_p(\varepsilon_n^2)$, so that $\|(\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}, (\hat{\mathbf{f}} - \bar{\mathbf{f}})/\mathbf{f})\|_\infty = O_p(\varepsilon_n^2)$.

Claim (i) follows from the uniform convergence of $\hat{f}$ to f that is assumed in the theorem: $\sup_{(x,y)\in S} |\hat{f}(x,y) - f(x,y)| = O_p(\varepsilon_n)$. Below, we give the proofs of Claims (ii) and (iii).

Proof of Claim (ii). For this claim we first prove that the map $\mathcal{G}'(\mathbf{0},\mathbf{0})$ is one-to-one. Suppose that $\mathcal{G}'(\mathbf{0},\mathbf{0})(\mathbf{d},\boldsymbol{\delta}) = \mathbf{0}$ for some $\mathbf{d} = (d_1,d_2,d_3)^\top$ and $\boldsymbol{\delta} = (\delta_1,\delta_2,\delta_3)^\top$. Then, by integrating the fourth component of $\mathcal{G}'(\mathbf{0},\mathbf{0})(\mathbf{d},\boldsymbol{\delta})$, we find that

$$0 = \int_S f(x,y)[\delta_1(x) + \delta_2(y) + \delta_3(m_J(x+y))]\,dx\,dy = d_1 \int_S f(x,y)\,dx\,dy,$$

where the first equation holds since the right hand side equals, up to sign change, the third component of $\mathcal{G}'(\mathbf{0},\mathbf{0})(\mathbf{d},\boldsymbol{\delta})$. Similarly, we get $d_2 = d_3 = 0$. Now, from $\mathcal{G}'(\mathbf{0},\mathbf{0})(\mathbf{0},\boldsymbol{\delta}) = \mathbf{0}$ we have

$$\begin{aligned} 0 &= \int_{S_1\times S_2\times S_3} (\mathbf{0}^\top, \boldsymbol{\delta}(x,y,z)^\top)\mathcal{G}'(\mathbf{0},\boldsymbol{\delta})(x,y,z)\,dx\,dy\,dz \\ &= -\int_S f(x,y)[\delta_1(x) + \delta_2(y) + \delta_3(m_J(x+y))]^2 dx\,dy. \end{aligned}$$

This implies

$$\delta_1(x) + \delta_2(y) + \delta_3(m_J(x+y)) = 0 \ \text{ a.e. on } S. \tag{A.5}$$

Arguing as in the proof of Theorem 2 using the last three equations of $\mathcal{G}'(\mathbf{0},\mathbf{0})(\mathbf{0},\boldsymbol{\delta}) = \mathbf{0}$, we obtain $\delta_j \equiv 0$ on S_j, $1 \le j \le 3$.

Next, we prove that the map $\mathcal{G}'(\mathbf{0},\mathbf{0})$ is onto. For a tuple $(\mathbf{c},\boldsymbol{\eta})$ with $\mathbf{c} = (c_1,c_2,c_3)^\top$ and $\boldsymbol{\eta}(x,y,z) = (\eta_1(x),\eta_2(y),\eta(z))^\top$, suppose that $\langle(\mathbf{c},\boldsymbol{\eta}), \mathcal{G}'(\mathbf{0},\mathbf{0})(\mathbf{d},\boldsymbol{\delta})\rangle = 0$ for all $(\mathbf{d},\boldsymbol{\delta}) \in \mathbb{R}^3 \times \mathcal{S}$. This implies

$$\begin{aligned}
0 &= \int_S f(x,y)\eta_1(x)\,dx\,dy, \\
0 &= \int_S f(x,y)\eta_2(y)\,dx\,dy, \\
0 &= \int_S f(x,y)\eta_3(m_J(x+y))\,dx\,dy, \\
0 &= \int_{J_2(x)} f(x,y)[\eta_1(x) + \eta_2(y) + \eta_3(m_J(x+y))]\,dy + c_1 f_1(x) + c_3 f_{w,1}(x), \\
0 &= \int_{J_1(y)} f(x,y)[\eta_1(x) + \eta_2(y) + \eta_3(m_J(x+y))]\,dx + c_2 f_2(y) + c_3 f_{w,2}(y), \\
0 &= \sum_{l=0}^{L(J)} \int_{J_{3l}(z)} f(x,(z+l)/J - x)[\eta_1(x) + \eta_2((z+l)/J - x) + \eta_3(z)]\,dx + c_3 f_{w,3}(z).
\end{aligned} \tag{A.6}$$

From the first three equations of (A.6), we get $c_1 + \vartheta c_3 = 0$ by integrating the fourth equation. Similarly, we obtain $c_2 + \vartheta c_3 = 0$ and $c_3 = 0$ by integrating the fifth and the sixth equations. This establishes $c_1 = c_2 = c_3 = 0$. Putting back these constant values to (A.6), multiplying $\eta_1(x)$, $\eta_2(y)$ and $\eta_3(z)$ to the right hand sides of the fourth, fifth and sixth equations, respectively, and then integrating them give

$$\int_S f(x,y)[\eta_1(x) + \eta_2(y) + \eta_3(m_J(x+y))]^2 \, dx \, dy = 0.$$

Going through the arguments in the proof of $\mathcal{G}'(\mathbf{0},\mathbf{0})$ being one-to-one and now using the first two equations of (A.6) give $\eta_1 = \eta_2 = \eta_3 \equiv 0$. Note that the first two equations can be written as $\int_{S_1} f_{w,1}(x)\eta_1(x)\,dx = 0$ and $\int_{S_2} f_{w,2}(y)\eta_2(y)\,dy = 0$, and thus in the latter proof $f_{w,j}$ for $j = 1, 2$ take the roles of f_j in the former proof. The foregoing arguments show that $(\mathbf{0},\mathbf{0})$ is the only tuple that is perpendicular to the range space of $\mathcal{G}'(\mathbf{0},\mathbf{0})$, which implies that $\mathcal{G}'(\mathbf{0},\mathbf{0})$ is onto.

To verify that the inverse map $\mathcal{G}'(\mathbf{0},\mathbf{0})^{-1}$ is bounded, it suffices to prove that the bijective linear operator $\mathcal{G}'(\mathbf{0},\mathbf{0})$ is bounded, owing to the bounded inverse theorem. Indeed, it holds that there exists a constant $C > 0$ such that $\|\mathcal{G}'(\mathbf{0},\mathbf{0})(\mathbf{d},\boldsymbol{\delta})\|_\infty \le C\|(\mathbf{d},\boldsymbol{\delta})\|_\infty$. This completes the proof of Claim (ii). $\square$

Proof of Claim (iii). We first note that $\hat{\mathcal{G}}'(\boldsymbol{\theta}_1,\mathbf{g}_1)(\mathbf{d},\boldsymbol{\delta}) - \hat{\mathcal{G}}'(\boldsymbol{\theta}_2,\mathbf{g}_2)(\mathbf{d},\boldsymbol{\delta}) = \mathcal{G}'(\boldsymbol{\theta}_1,\mathbf{g}_1)(\mathbf{d},\boldsymbol{\delta}) - \mathcal{G}'(\boldsymbol{\theta}_2,\mathbf{g}_2)(\mathbf{d},\boldsymbol{\delta})$. From this, we get that, for each given $r > 0$

$$\|\hat{\mathcal{G}}'(\boldsymbol{\theta}_1,\mathbf{g}_1)(\mathbf{d},\boldsymbol{\delta}) - \hat{\mathcal{G}}'(\boldsymbol{\theta}_2,\mathbf{g}_2)(\mathbf{d},\boldsymbol{\delta})\|_\infty \le 6\,(1+r) \max_{1\le j\le 3} \sup_{u\in S_j} f_{w,j}(u) \|\mathbf{g}_2 - \mathbf{g}_1\|_\infty$$

for all $(\boldsymbol{\theta}_1,\mathbf{g}_1)$, $(\boldsymbol{\theta}_2,\mathbf{g}_2) \in B_r(\mathbf{0},\mathbf{0})$ and for all $(\mathbf{d},\boldsymbol{\delta})$ with $\|(\mathbf{d},\boldsymbol{\delta})\|_\infty = 1$. For this, we used the inequality

$$\sup_{(x,y,z)\in S_1\times S_2\times S_3} |\kappa(x,y,z;\mathbf{g}_2,\boldsymbol{\delta}) - \kappa(x,y,z;\mathbf{g}_1,\boldsymbol{\delta})| \le 3\,\|\boldsymbol{\delta}\|_\infty (2 + \|\mathbf{g}_1\|_\infty + \|\mathbf{g}_2\|_\infty)\|\mathbf{g}_2 - \mathbf{g}_1\|_\infty.$$

This completes the proof of (iii).

A.4 Proof of Theorem 4

Let $\hat{f}^A(x,y)$ be the first entry of $\hat{\boldsymbol{\eta}}^A(x,y)$, where $\hat{\boldsymbol{\eta}}^A$ is defined as $\hat{\boldsymbol{\eta}}$ at (3.5) with $\hat{\mathbf{b}}$ being replaced by $\hat{\mathbf{b}} - E\hat{\mathbf{b}}$. Likewise, define $\hat{f}^B(x,y)$ with $\hat{\mathbf{b}}(x,y)$ being replaced by $E\hat{\mathbf{b}}(x,y) -$

$(f(x,y), h_1\partial f(x,y)/\partial x, h_2\partial f(x,y)/\partial y)^\top$. Then, $\hat{f}(x,y) = f(x,y) + \hat{f}^A(x,y) + \hat{f}^B(x,y)$. Define $\hat{\boldsymbol{\mu}}^A$ and $\hat{\boldsymbol{\mu}}^B$ as $\hat{\boldsymbol{\mu}}$ at (4.2) with $\hat{f} - f$ being replaced by $\hat{f}^A$ and $\hat{f}^B$, respectively, and $\bar{\mathbf{f}}^s/\mathbf{f} = (\bar{f}_1^s/f_1, \bar{f}_2^s/f_2, \bar{f}_3^s/f_3)$ along with $\bar{\boldsymbol{\theta}}^s - \mathbf{1} = (\bar{\theta}_1^s - 1, \bar{\theta}_2^s - 1, \bar{\theta}_3^s - 1)$ for $s = A$ and B as the solution of the backfitting equation (4.4) with $\hat{\boldsymbol{\mu}}$ being replaced by $\hat{\boldsymbol{\mu}}^s$, subject to the constraints (4.5). Since the backfitting equation (4.4) is linear in $\hat{\boldsymbol{\mu}}$, we get that $\bar{\mathbf{f}} = \mathbf{f} + \bar{\mathbf{f}}^A + \bar{\mathbf{f}}^B$ and $\bar{\boldsymbol{\theta}} = \bar{\boldsymbol{\theta}}^A - \mathbf{1} + \bar{\boldsymbol{\theta}}^B$.

For simplicity, write the backfitting equation (4.4) as $\boldsymbol{\delta} = \mathbf{d} + \hat{\boldsymbol{\mu}} - \mathbf{T}\boldsymbol{\delta}$ with an appropriate definition of the linear operator $\mathbf{T}$. From the definitions of $\bar{\mathbf{f}}^A$ and $\bar{\boldsymbol{\theta}}^A$ we have $\bar{\mathbf{f}}^A/\mathbf{f} = \bar{\boldsymbol{\theta}}^A - \mathbf{1} + \hat{\boldsymbol{\mu}}^A - \mathbf{T}(\bar{\mathbf{f}}^A/\mathbf{f})$. From Lemma 2 below, we obtain

$$\bar{\mathbf{f}}^A/\mathbf{f} - \hat{\boldsymbol{\mu}}^A = \bar{\boldsymbol{\theta}}^A - \mathbf{1} - \mathbf{T}(\bar{\mathbf{f}}^A/\mathbf{f} - \hat{\boldsymbol{\mu}}^A) + o_p(n^{-2/5})$$

uniformly on $S_1 \times S_2 \times S_3$. This implies $\bar{\mathbf{f}}^A/\mathbf{f} - \hat{\boldsymbol{\mu}}^A = o_p(n^{-2/5})$ uniformly on $S_1 \times S_2 \times S_3$ and $\bar{\boldsymbol{\theta}}^A - \mathbf{1} = o_p(n^{-2/5})$.

Now, for the deterministic part $\bar{\mathbf{f}}^B$, recall the definitions of $\tilde{f}^B$ and $\tilde{\boldsymbol{\mu}}^B$ at (4.6) and thereafter, respectively. Let $\mathbf{r}_n = \hat{\boldsymbol{\mu}}^B - n^{-2/5}\tilde{\boldsymbol{\mu}}^B$. According to Lemma 2, $\mathbf{r}_n = o(n^{-2/5})$ on $S_1' \times S_2' \times S_3'$, where S_j' is a subset of S_j with the property that $\text{mes}(S_j - S_j') = O(n^{-1/5})$. We also get $\mathbf{r}_n = O(n^{-2/5})$ on $S_1 \times S_2 \times S_3$. This implies $\mathbf{T}(\mathbf{r}_n) = o(n^{-2/5})$, so that

$$\bar{\mathbf{f}}^B/\mathbf{f} - \mathbf{r}_n = \bar{\boldsymbol{\theta}}^B - \mathbf{1} + n^{-2/5}\tilde{\boldsymbol{\mu}}^B - \mathbf{T}(\bar{\mathbf{f}}^B/\mathbf{f} - \mathbf{r}_n) + o_p(n^{-2/5})$$

uniformly on $S_1 \times S_2 \times S_3$. Thus, $(\bar{\mathbf{f}}^B/\mathbf{f}, \bar{\boldsymbol{\theta}}^B - \mathbf{1})$ equals the solution of the backfitting equation $\boldsymbol{\delta} = \mathbf{d} + n^{-2/5}\tilde{\boldsymbol{\mu}}^B - \mathbf{T}\boldsymbol{\delta}$, up to an additive term whose jth component has a magnitude of an order $o(n^{-2/5})$ on S_j' and $O(n^{-2/5})$ on the whole set S_j.

The asymptotic distribution of $\big((\bar{f}_j(u_j) - f_j(u_j))/f_j(u_j) : 1 \le j \le 3\big)$ for fixed $u_j \in S_{j,c} \cap S_j^{\text{o}}$ is then readily obtained from the above results. The asymptotic mean is given as the solution $(\delta_j(u_j) : 1 \le j \le 3)$ of the backfitting equation (4.4) with $\hat{\mu}_j$ being replaced by $n^{-2/5}\tilde{\mu}_j^B$, subject to the constraint (4.5). The asymptotic variances are derived from

those of $\tilde{\mu}_j^A$, where

$$\tilde{\mu}_1^A(x) = f_{w,1}(x)^{-1} \int_{J_2(x)} \tilde{f}^A(x,y)\, dy,$$
$$\tilde{\mu}_2^A(y) = f_{w,2}(y)^{-1} \int_{J_1(y)} \tilde{f}^A(x,y)\, dx,$$
$$\tilde{\mu}_3^A(z) = f_{w,3}(z)^{-1} \sum_{l=0}^{L(J)} \int_{J_{3l}(z)} \tilde{f}^A(x,(z+l)/J - x)\, dx$$

and $\tilde{f}^A(x,y) = n^{-1}\sum_{i=1}^n [K_{h_1}(X_i - x)K_{h_2}(Y_i - y)W_i - E\left(K_{h_1}(X_i - x)K_{h_2}(Y_i - y)W_i\right)]$. This is due to (A.9), (A.10) and the corresponding property for $\hat{\mu}_3^A$ in the proof of Lemma 2 below.

To compute $\mathrm{var}(\tilde{\mu}_1^A(u_1))$, we note that, due to the assumption (A7) and thus from Lemma 1, we may find constants $C > 0$ and $\alpha > 1/2$ such that $J_2^o(u; Ch_1^\alpha + h_2) \subset J_2^o(u_1; h_2)$ for all u with $|u - u_1| \le h_1$, if n is sufficiently large. Note that $J_2^o(u; Ch_1^\alpha + h_2)$ is inside $J_2^o(u; h_2)$ at a depth Ch_1^α. Then, it can be shown that, for all (u, v) with $|u - u_1| \le h_1$ and $v \in J_2^o(u; Ch_1^\alpha + h_2)$, the set $\{(v - y)/h_2 : y \in J_2(u_1)\}$ covers the interval $[-1, 1]$, the support of the kernel K. This implies that $K_{h_1}(u - u_1)\nu(u_1, v) = K_{h_1}(u - u_1)$ for all (u, v) with $|u - u_1| \le h_1$ and $v \in J_2^o(u; Ch_1^\alpha + h_2)$, where $\nu(u_1, v) = \int_{J_2(u_1)} K_{h_2}(v - y)\, dy$. Using this and the fact that the Lebesgue measure of the set difference $J_2(u) - J_2^o(u; Ch_1^\alpha + h_2)$ has a magnitude of order $n^{-\min\{1,\alpha\}/5}$, we get

$$\begin{aligned}
\mathrm{var}(\tilde{\mu}_1^A(u_1)) &= f_{w,1}(u_1)^{-2} n^{-1} h_1^{-1} \int_S \frac{1}{h_1} K\left(\frac{u - u_1}{h_1}\right)^2 \nu(u_1, v)^2 f(u, v)\, du\, dv + O(n^{-1}) \\
&= f_{w,1}(u_1)^{-2} n^{-1} h_1^{-1} \int_{|u-u_1| \le h_1} \int_{J_2^o(u; Ch_1^\alpha + h_2)} \frac{1}{h_1} K\left(\frac{u - u_1}{h_1}\right)^2 \nu(u_1, v)^2 \\
&\qquad \times f(u, v)\, dv\, du + o(n^{-1}h^{-1}) \\
&= f_{w,1}(u_1)^{-2} n^{-1} h_1^{-1} \int_S \frac{1}{h_1} K\left(\frac{u - u_1}{h_1}\right)^2 f(u, v)\, du\, dv + o(n^{-1}h^{-1}) \\
&= n^{-1} h_1^{-1} f_{w,1}(u_1)^{-1} \int K^2(u)\, du + o(n^{-1}h^{-1}).
\end{aligned}$$

The last equation holds since $u_1 \in S_{1,c}$, so that $f_{w,1}$ is continuous at u_1, and it is a fixed point in the interior of S_1. Similarly, we obtain

$$\mathrm{var}(\tilde{\mu}_2^A(u_2)) = n^{-1} h_2^{-1} f_{w,2}(u_2)^{-1} \int K^2(u)\, du + o(n^{-1}h^{-1}).$$

The calculation of the asymptotic variance of $\tilde{\mu}_3^A(u_3)$ is more involved than those of $\mathrm{var}(\tilde{\mu}_j^A(u_j))$ for $j = 1, 2$. For this, we observe that, if $l \neq l'$, then for any given $z \in [0,1]$ and $(u,v) \in \mathcal{I}$ we have

$$\begin{aligned}&\pi_{l,l'}(z,u,v,x,x') \\ &\quad \equiv K_{h_1}(u-x)K_{h_2}\left(v - \frac{z+l}{J} + x\right) K_{h_1}(u-x')K_{h_2}\left(v - \frac{z+l'}{J} + x'\right) = 0\end{aligned}$$

for all x, x' except the case $(z+l)/J - x = (z+l')/J - x'$, if n is sufficiently large. This implies that

$$\begin{aligned}\mathrm{var}(\tilde{\mu}_3^A(u_3)) = f_{w,3}(u_3)^{-2} n^{-1} \sum_{l=0}^{L(J)} \int_{J_{3l}(u_3)} \int_{J_{3l}(u_3)} \int_S \pi_l(u_3,u,v,x,x') f(u,v)\, du\, dv\, dx\, dx' \\ + O(n^{-1}),\end{aligned}$$

where $\pi_l = \pi_{l,l}$. From Lemma 1 again, we may find constants $C > 0$ and $\alpha > 1/2$ such that $J_2^{\mathrm{o}}(x; Ch_1^\alpha + h_2) \subset J_2^{\mathrm{o}}(u; h_2)$ for all $x, u \in (a_{k-1}^1, a_k^1) \cap S_1$ with $|u-x| \leq h_1$, $1 \leq k \leq L_1$. Define a subset $J'_{3l}(u_3)$ of $[0,1]$ such that $x \in J'_{3l}(u_3)$ if and only if $x \in J_{3l}(u_3 + J(h_2 + Ch_1^\alpha)t)$ for all $t \in [-1,1]$. Then, for a given $u \in S_{1,c}$, it follows that

$$[-1,1] \subset \left\{ \frac{v - (u_3 + l)/J + x}{h_2} : v \in J_2(u) \right\}$$

for all $x \in J'_{3l}(u_3)$ such that $|x-u| \leq h_1$ and x lies in the same partition (a_{k-1}^1, a_k^1) as u. This holds since $x \in J_{3l}(z)$ implies $(z+l)/J - x \in J_2(x)$. This entails that, for $x \in J'_{3l}(u_3) \cap S_{1,c}^{\mathrm{o}}(h_1)$,

$$\begin{aligned}&\int_S \pi_l(u_3,u,v,x,x')\, du\, dv \\ &\quad = \int_{[-1,1]^2} K(t)K(s)h_1^{-1}K\left(t + \frac{x-x'}{h_1}\right) h_2^{-1} K\left(s + \frac{x'-x}{h_2}\right) dt\, ds \\ &\quad = (K*K)_{h_1}(x-x')(K*K)_{h_2}(x-x'),\end{aligned}$$

where $K * K$ denotes the convolution of K defined by $K * K(u) = \int K(t)K(t+u)\, dt$. Here and below, $S_{j,c}^{\mathrm{o}}(h)$ for a small number $h > 0$ denotes the set of $x \in S_{j,c}$ such that $x + ht$ belongs to $S_{j,c}$ for all $t \in [-1,1]$.

Because of the assumption (A7) and the fact that u_3 is a fixed point in $S_{3,c}$, we get that $\sum_{l=0}^{L(J)} \mathrm{mes}[J_{3l}(u_3) \triangle J'_{3l}(u_3)]$ is of order $o(1)$. This and the foregoing arguments give

$$\begin{aligned}
\mathrm{var}(\tilde{\mu}_3^A(u_3)) &= f_{w,3}(u_3)^{-2} n^{-1} \sum_{l=0}^{L(J)} \int_{J_{3l}(u_3)} \int_{J'_{3l}(u_3) \cap S^{\mathrm{o}}_{1,c}(h_1)} \int_S \pi_l(u_3, u, v, x, x')\, du\, dv \\
&\qquad \times f\left(x, \frac{u_3 + l}{J} - x\right) dx\, dx' + o(n^{-4/5}) \\
&= f_{w,3}(u_3)^{-2} n^{-1} \sum_{l=0}^{L(J)} \int_{J_{3l}(u_3)} \int_{J_{3l}(u_3)} (K * K)_{h_1}(x - x')(K * K)_{h_2}(x - x') \\
&\qquad \times f\left(x, \frac{u_3 + l}{J} - x\right) dx\, dx' + o(n^{-4/5}).
\end{aligned}$$

Let $J^{\mathrm{o}}_{3l}(u_3; 2h_1)$ denote a subset of $J_{3l}(u_3)$ such that $x \in J^{\mathrm{o}}_{3l}(u_3; 2h_1)$ if and only if $x - 2h_1 t \in J_{3l}(u_3)$ for all $t \in [-1, 1]$. Then,

$$\begin{aligned}
&\sum_{l=0}^{L(J)} \int_{J_{3l}(u_3)} \int_{J_{3l}(u_3)} (K * K)_{h_1}(x - x')(K * K)_{h_2}(x - x') f\left(x, \frac{u_3 + l}{J} - x\right) dx'\, dx \\
&= h_2^{-1} \sum_{l=0}^{L(J)} \int_{J^{\mathrm{o}}_{3l}(u_3; 2h_1)} f\left(x, \frac{u_3 + l}{J} - x\right) dx \int_{-2}^{2} [K * K(t)][K * K(h_1 t / h_2)]\, dt + O(1) \\
&= h_2^{-1} \sum_{l=0}^{L(J)} \int_{J_{3l}(u_3)} f\left(x, \frac{u_3 + l}{J} - x\right) dx \int_{-2}^{2} [K * K(t)][K * K(h_1 t / h_2)]\, dt + O(1) \\
&= h_2^{-1} f_{w,3}(u_3) \int_{-2}^{2} [K * K(t)][K * K(h_1 t / h_2)]\, dt + O(1)
\end{aligned}$$

This with Lemma 3 below completes the proof of Theorem 4.

LEMMA 1 *Under the condition (A7) with the constants $C > 0$ and $\alpha > 1/2$, it follows that (i) $J^{\mathrm{o}}_2(u_1 : Ch_1^\alpha + h_2) \subset J^{\mathrm{o}}_2(u_2; h_2)$ for any $u_1, u_2 \in (a^1_{k-1}, a^1_k) \cap S_1$ with $|u_1 - u_2| \le h_1$, $1 \le k \le L_1$; (ii) $J^{\mathrm{o}}_1(u_1 : Ch_2^\alpha + h_1) \subset J^{\mathrm{o}}_1(u_2; h_1)$ for any $u_1, u_2 \in (a^2_{k-1}, a^2_k) \cap S_2$ with $|u_1 - u_2| \le h_2$, $1 \le k \le L_2$.*

Proof of Lemma 1. We apply (A7) to the choice $\epsilon_n = h_1$. Suppose a point $y \in J^{\mathrm{o}}_2(u_1; Ch_1^\alpha + h_2)$. This implies $y + h_2 t + Ch_1^\alpha s \in J_2(u_1)$ for all $s, t \in [-1, 1]$. This holds since $|(h_2 t + Ch_1^\alpha s)/(h_2 + Ch_1^\alpha)| \le 1$ for all $s, t \in [-1, 1]$. By (A7), $y + h_2 t \in J^{\mathrm{o}}_2(u_1; Ch_1^\alpha) \subset J_2(u_2)$ for all $t \in [-1, 1]$, so that we get $y \in J^{\mathrm{o}}_2(u_2; h_2)$. The proof of (ii) is the same. □

LEMMA 2 *Under the conditions of Theorem 4, It follows that* $\mathbf{T}\hat{\boldsymbol{\mu}}^A = o_p(n^{-2/5})$ *uniformly on* $S_1 \times S_2 \times S_3$. *Furthermore,* $\hat{\boldsymbol{\mu}}^B = n^{-2/5}\tilde{\boldsymbol{\mu}}^B + o(n^{-2/5})$ *uniformly on* $S^{\mathrm{o}}_{1,c}(h_1) \times S^{\mathrm{o}}_{2,c}(h_2) \times S^{\mathrm{o}}_{3,c}(C'n^{-\min\{1,\alpha\}/5})$ *for a sufficiently large* $C' > 0$, *and* $\hat{\boldsymbol{\mu}}^B(u) = n^{-2/5}\tilde{\boldsymbol{\mu}}^B(u) + O(n^{-2/5})$ *uniformly uniformly on* $S_1 \times S_2 \times S_3$.

Proof of Lemma 2. From the standard theory of kernel smoothing it follows that

$$\sup_{(x,y)\in S} |\hat{f}^A(x,y)| = O_p(n^{-3/10}\sqrt{\log n}). \tag{A.7}$$

Also, we have $\mathbf{A}(x,y) = \mathrm{diag}(1,\nu_2,\nu_2)$ for all (x,y) with $x \in S^{\mathrm{o}}_{1,c}(h_1)$ and $y \in J^{\mathrm{o}}_2(x; Ch_1^\alpha + h_2)$, where $C > 0$ and $\alpha > 1/2$ are the constants in Assumption (A7) and $\nu_2 = \int u^2 K(u)\,du$. Define $\mathcal{J} = \{(x,y) \in S : x \in S^{\mathrm{o}}_{1,c}(h_1),\, y \in J^{\mathrm{o}}_2(x; Ch_1 + h_2)\}$. From the simplification of $\mathbf{A}(x,y)$ on $\mathcal{J}$, we get

$$\hat{f}^A(x,y) = \tilde{f}^A(x,y), \quad (x,y) \in \mathcal{J}. \tag{A.8}$$

From (A.7) and (A.8) we have

$$\hat{\mu}_1^A(x) = \tilde{\mu}_1^A(x) + O_p(n^{-(3+2r)/10}\sqrt{\log n}) \text{ uniformly for } x \in S^{\mathrm{o}}_{1,c}(h_1), \tag{A.9}$$

where $r = \min\{1,\alpha\}$. Note that $r > 1/2$. Similarly, we get

$$\hat{\mu}_2^A(y) = \tilde{\mu}_2^A(y) + O_p(n^{-(3+2r)/10}\sqrt{\log n}) \text{ uniformly for } y \in S^{\mathrm{o}}_{2,c}(h_2). \tag{A.10}$$

For the treatment of $\hat{\mu}_3^A$, we first note that $\mathbf{A}(x,(z+l)/J - x) = \mathrm{diag}(1,\nu_2,\nu_2)$ for all $x \in J'_{3l}(z) \cap S^{\mathrm{o}}_{1,c}(h_1)$, where the set $J'_{3l}(z)$ is defined in the proof of Theorem 4. In fact,

$$(x,(z+l)/J - x) \in \mathcal{J} \text{ if and only if } x \in J'_{3l}(z) \cap S^{\mathrm{o}}_{1,c}(h_1). \tag{A.11}$$

This implies that, for all $0 \le l \le L(J)$,

$$\hat{f}^A\left(x, \frac{z+l}{J} - x\right) = \tilde{f}^A\left(x, \frac{z+l}{J} - x\right), \quad x \in J'_{3l}(z) \cap S^{\mathrm{o}}_{1,c}(h_1). \tag{A.12}$$

Due to the condition (A7) we can take a constant $C' > 0$ such that, uniformly for $z \in S^{\mathrm{o}}_{3,c}(C'n^{-r/5})$, we have $\sum_{l=0}^{L(J)} \mathrm{mes}[J_{3l}(z) \triangle J'_{3l}(z)] = O(n^{-r/5})$. Then, from (A.7) and (A.12)

we have

$$\begin{aligned}
&\sum_{l=0}^{L(J)} \int_{J_{3l}(z)} \hat{f}^A(x,(z+l)/J-x)\,dx \\
&\quad = \sum_{l=0}^{L(J)} \int_{J'_{3l}(z)\cap S^{\mathrm{o}}_{1,c}(h_1)} \tilde{f}^A(x,(z+l)/J-x)\,dx \\
&\qquad\qquad + O_p(n^{-3/10}\sqrt{\log n}) \sum_{l=0}^{L(J)} \mathrm{mes}[J_{3l}(z) \triangle (J'_{3l}(z)\cap S^{\mathrm{o}}_{1,c}(h_1))] \\
&\quad = \sum_{l=0}^{L(J)} \int_{J_{3l}(z)} \tilde{f}^A(x,(z+l)/J-x)\,dx + o_p(n^{-2/5})
\end{aligned}$$

uniformly for $z \in S^{\mathrm{o}}_{3,c}(C'n^{-r/5})$. This implies $\hat{\mu}_3^A(z) = \tilde{\mu}_3^A(z) + o_p(n^{-2/5})$ uniformly for $z \in S^{\mathrm{o}}_{3,c}(C'n^{-r/5})$. This together with (A.9), (A.10) and Lemma 3 gives $\mathbf{T}\hat{\boldsymbol{\mu}}^A = o_p(n^{-2/5})$ uniformly on $S_1 \times S_2 \times S_3$, since $\mathbf{T}\tilde{\boldsymbol{\mu}}^A = o_p(n^{-2/5})$ uniformly on the set and the Lebesgue measures of the set differences $S_1 - S^{\mathrm{o}}_{1,c}(h_1)$ and $S_2 - S^{\mathrm{o}}_{2,c}(h_2)$ are of order $n^{-1/5}$ and that of $S_3 - S^{\mathrm{o}}_{3,c}(C'n^{-r/5})$ is of order $n^{-r/5}$.

To prove the second part of the lemma, recall that $\mathbf{A}(x,y) = \mathrm{diag}(1,\nu_2,\nu_2)$ on $\mathcal{J}$. In fact, for $(x,y) \in \mathcal{J}$

$$\int_S \left(\frac{u-x}{h_1}\right)^j \left(\frac{v-y}{h_2}\right)^k K_{h_1}(u-x)K_{h_2}(v-y)\,du\,dv = 0$$

whenever j or k is an odd integer. This implies $\hat{f}^B(x,y) = n^{-2/5}\tilde{f}^B(x,y) + o(n^{-2/5})$ uniformly for $(x,y) \in \mathcal{J}$. We also get $\hat{f}^B(x,y) = O(n^{-2/5})$ uniformly for $(x,y) \in S$. We apply the same arguments as in the proof of the first part, to obtain

$$\begin{aligned}
\hat{\mu}_1^B(x) &= n^{-2/5}\tilde{\mu}_1^B(x) + o(n^{-2/5}) \ \text{ uniformly for } \ x \in S^{\mathrm{o}}_{1,c}(h_1), \\
\hat{\mu}_2^B(y) &= n^{-2/5}\tilde{\mu}_2^B(y) + o(n^{-2/5}) \ \text{ uniformly for } \ y \in S^{\mathrm{o}}_{2,c}(h_2).
\end{aligned}$$

From (A.11) it follows that

$$\hat{f}^B\left(x, \frac{z+l}{J} - x\right) = n^{-2/5}\,\tilde{f}^B\left(x, \frac{z+l}{J} - x\right) + o(n^{-2/5}).$$

for all (x,z) such that $x \in J'_{3l}(z) \cap S^{\mathrm{o}}_{1,c}(h_1)$ and $z \in S_3$. From this and the fact that $\sum_{l=0}^{L(J)} \mathrm{mes}[J_{3l}(z) \triangle J'_{3l}(z)] = o(1)$ uniformly for $z \in S^{\mathrm{o}}_{3,c}(C'n^{-r/5})$, we obtain

$$\hat{\mu}_3^B(z) = n^{-2/5}\tilde{\mu}_3^B(z) + o(n^{-2/5}) \ \text{ uniformly for } \ z \in S^{\mathrm{o}}_{3,c}(C'n^{-r/5}),$$

where C' is the constant C' in the proof of the first part. This completes the proof of the lemma. $\square$

LEMMA 3 *Under the conditions of Theorem 4, it follows that*

$$\sup_{u\in S_j} |\hat{\mu}_j^A(u)| = O_p(n^{-2/5}\sqrt{\log n}), \quad 1 \le j \le 3.$$

Proof of Lemma 3. We give the proof for $\hat{\mu}_1^A$ only. The others are similar. For (x, y) with $x \in S_1$ and $y \in J_2^o(x; Ch_1^\alpha + h_2)$, we have

$$\hat{f}^A(x,y) = \varphi_1(x)\hat{a}_1(x,y) + \varphi_2(x)\hat{a}_2(x,y) + \varphi_3(x)\hat{a}_3(x,y),$$

where φ_j for $j = 1, 2, 3$ are some bounded functions, $\hat{a}_1 = \hat{b}_{00}$, $\hat{a}_2 = \hat{b}_{10}$ and $\hat{a}_3 = \hat{b}_{01}$ with

$$\begin{aligned}\hat{b}_{jk}(x,y) =& n^{-1}\sum_{i=1}^n \left[\left(\frac{X_i - x}{h_1}\right)^j \left(\frac{Y_i - y}{h_2}\right)^k K_{h_1}(X_i - x)K_{h_2}(Y_i - y)W_i \right.\\ & \left. - \; E\left(\frac{X_i - x}{h_1}\right)^j \left(\frac{Y_i - y}{h_2}\right)^k K_{h_1}(X_i - x)K_{h_2}(Y_i - y)W_i\right]\end{aligned}$$

The lemma follows from (A.5) and using

$$\begin{aligned}&\sup_{x\in S_1} \text{mes}[J_2(x) - J_2^o(x; Ch_1^\alpha + h_2)] = O_p(n^{-r/5}),\\ &\sup_{x\in S_1} \Big| \int_{J_2(x)} \hat{a}_j(x,y)\, dy\Big| = O_p(n^{-2/5}\sqrt{\log n}), \quad 1 \le j \le 3. \;\square\end{aligned}$$

References

Cheng, M.-Y. (1997). A bandwidth selector for local linear density estimators, *Annals of Statistics*, 25, 1001–1013.

Guillot, D., Khare, A. and Rajaratnam, B. (2013). Classification of measurable solutions of Cauchy's functional equations, and operators satisfying the Chain Rule. Preprint arXiv:1312.6297 [math.FA]

Jiang, J., Fan, Y. and Fan, J. (2010). Estimation in additive models with highly correlated covariates, *Annals of Statistics*, 38, 1403–1432.

Kuang, D., Nielsen, B. and Nielsen, J.P. (2008). Identification of the age-period-cohort model and the extended chain-ladder model, *Biometrika*, 95, 979–986.

Kuang, D., Nielsen, B. and Nielsen, J.P. (2009). Chain-Ladder as Maximum Likelihood Revisited, *Annals of Actuarial Science*, 4, 105–121.

Lee, Y.K., Mammen, E. and Park, B.U. (2010). Backfitting and smooth backfitting for additive quantile models, *Annals of Statistics*, 38, 2857–2883.

Lee, Y.K., Mammen, E. and Park, B.U. (2012). Flexible generalized varying coefficient regression models, *Annals of Statistics*, 40, 1906–1933.

Lee, Y.K., Mammen, E. and Park, B.U. (2013). Backfitting and smooth backfitting in varying coefficient quantile regression, *Econometrics Journal*, in print.

Linton, O. and Nielsen, J.P. (1995). A kernel method of estimating structured nonparametric regression, *Biometrika*, 82, 93–100.

Mammen, E., Linton, O. and Nielsen, J.P. (1999). The existence and asymptotic properties of a backfitting algorithm under weak conditions, *Annals of Statistics*, 27, 1443–1490.

Mammen, E. and Nielsen, J.P. (2003). Generalised structured Models, *Biometrika*, 90, 551–566.

Mammen, E., Martínez-Miranda, M.D. and Nielsen, J.P. (2013). Structured density forecasting applied to non-life insurance and mesothelioma mortality. *Submitted*.

Mammen, E. and Park, B.U. (2005). Bandwidth selection for smooth backfitting in additive models, *Annals of Statistics*, 33, 1260–1294.

Mammen, E. and Park, B.U. (2006). A simple smooth backfitting method for additive models, *Annals of Statistics*, 34, 2252–2271.

Martínez-Miranda, M.D., Nielsen B., Nielsen J.P. and Verrall, R.J. (2011). Cash flow simulation for a model of outstanding liabilities based on claim amounts and claim numbers, *Astin Bulletin*, 41, 107–129.

Martínez-Miranda, M.D., Nielsen, J.P., Sperlich, S. and Verrall, R.J. (2013). Continuous Chain Ladder: Reformulating and generalising a classical insurance problem, *Expert Systems with Applications*, 40, 5588–5603.

Martínez-Miranda, M.D., Nielsen,J.P. and Verrall, R.J. (2012). Double Chain Ladder, *Astin Bulletin*, 42, 59–76.

Nielsen, J.P. (1999). Multivariate boundary kernels from local linear estimation, *Scandinavian Actuarial Journal*, 1, 93–95.

Nielsen, J.P. and Linton, O. (1998). An optimization interpretation of integration and backfitting estimators for separable nonparametric models, *Journal of Royal Statistical Society*, Series B, 60, 217–222.

Nielsen, J.P. and Sperlich, S. (2005). Smooth backfitting in practice, *Journal of Royal Statistical Society*, Series B, 67, 43–61.

Opsomer, J.D. and Ruppert, D. (1997). Fitting a bivariate additive model by local polynomial regression, *Annals of Statistics*, 25, 186–211.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

Verrall, R.J., Nielsen, J.P. and Jessen, A. (2010). Prediction of RBNS and IBNR claims using claim amounts and claim counts, *Astin Bulletin*, 40, 871–887.

Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing*, Chapman and Hall, London.

Yu, K., Park, B.U. and Mammen, E. (2008). Smooth backfitting in generalized additive models, *Annals of Statistics*, 36, 228–260.

Zhang, X., Park, B.U. and Wang, J.-L. (2013). Time-varying additive models for longitudinal data, *Journal of the American Statistical Association*, 108, 983–998.